APLICACIÓN DE MODELOS PREDICTIVOS EN LA CONTINUIDAD
DEL PROCESO DE FORMACIÓN DE ESTUDIANTES DE PREGRADO EN LA PONTIFICIA
UNIVERSIDAD JAVERIANA CALI

JUAN FELIPE MOSQUERA GARCIA

Nota de Aceptación

Certificamos que el presente Trabajo de Grado
Satisface, en alcances y calidad, todos los requisitos
que demanda un Trabajo de Grado de Maestría.

MARIA CONSTANZA PABÓN
Director

MONICA VIVIANA RODRIGUEZ
Jurado

GERARDO MAURICIO SARRIA
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la
Pontificia Universidad Javeriana Cali, para optar el título de
Magister en Ingeniería de Software.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 02 de septiembre de 2022

Acta de Correcciones al Documento de Trabajo de Grado

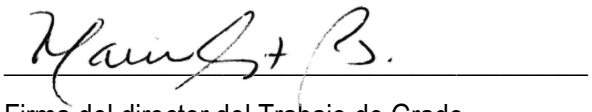
Santiago de Cali, 1 de septiembre de 2022

Autor: Juan Felipe Mosquera García

Título del Trabajo de Grado: "APLICACIÓN DE MODELOS PREDICTIVOS EN LA CONTINUIDAD DEL PROCESO DE FORMACIÓN DE ESTUDIANTES DE PREGRADO EN LA PONTIFICIA UNIVERSIDAD JAVERIANA CALI"

Director: María Constanza Pabón

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.



Firma del director del Trabajo de Grado

DATOS DEL ESTUDIANTE

TRABAJO DE GRADO

1. NOMBRE COMPLETO: Juan Felipe Mosquera García
2. DIRECCIÓN: Carrera 121A 42-16
3. TELÉFONOS DE CONTACTO: 3176585260
4. CORREO ELECTRÓNICO: juan.mosquera@javerianacali.edu.co
5. PROFESIÓN: Ingeniero Informático
6. UNIVERSIDAD: Universidad Autónoma de Occidente
7. EMPRESA: Pragma S.A.
8. CARGO: Desarrollador de Software

RESUMEN

TRABAJO DE GRADO

La minería de datos educativos es una nueva tendencia en el campo de la minería de datos y el descubrimiento de conocimientos en bases de datos que se centra en la minería de patrones útiles y el descubrimiento de conocimientos útiles de los sistemas de información educativos. A partir de la aplicación de técnicas de predicción y clasificación se pretende analizar la información recopilada de los estudiantes a través de sus años de estudio y proporcionar clasificaciones basadas en datos recopilados para predecir y clasificar a aquellos estudiantes que pueden continuar su proceso y/o transición educativa desde sus años de estudio en pregrado hacia los diferentes posgrados de la Pontificia Universidad Javeriana Cali. El enfoque del proyecto está dirigido a predecir con algunas características de los estudiantes aplicando la metodología CRISP-DM para los pronósticos de los algoritmos de clasificación resultantes.

ABSTRACT

TRABAJO DE GRADO

Educational data mining is a new trend in the field of data mining and knowledge discovery in databases that focuses on mining useful patterns and discovering useful knowledge from educational information systems. From the application of prediction and classification techniques it is intended to analyze the information collected from students throughout their years of study and provide classifications based on data collected to predict and classify those students who can continue their process and / or transition educational from his years of undergraduate study to the different postgraduate degrees of the Pontificia Universidad Javeriana Cali. The focus of the project will be aimed at predicting or classifying with those predominant characteristics of the students applying the CRISP-DM methodology for the predictions of the resulting classification algorithms.

Pontificia Universidad
JAVERIANA
Cali

APLICACIÓN DE MODELOS PREDICTIVOS EN LA CONTINUIDAD DEL PROCESO DE FORMACIÓN DE ESTUDIANTES DE PREGRADO EN LA PONTIFICIA UNIVERSIDAD JAVERIANA CALI

Juan Felipe Mosquera García
Código 00020401701

Proyecto de trabajo de grado para optar al título de
Magister en Ingeniería de Software

Director(a)
Dra. María Constanza Pabón

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN INGENIERÍA DE SOFTWARE
SANTIAGO DE CALI, JUNIO 21 DE 2021

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

La minería de datos educativos (EDM) es una nueva tendencia en el campo de la minería de datos y el descubrimiento de conocimientos en bases de datos (KDD) que se centra en la minería de patrones útiles y el descubrimiento de conocimientos útiles de los sistemas de información educativos, como sistemas de admisión, sistemas de registro, gestión de cursos, sistemas de aprendizaje (Moodle, Blackboard, etc…), y cualquier otro sistema que atienda a estudiantes de diferentes niveles educativos, desde escuelas y colegios hasta universidades. Los investigadores en este campo se centran en descubrir conocimientos útiles, ya sea para ayudar a los institutos educativos a gestionar mejor a sus estudiantes, o para ayudar a los estudiantes a gestionar mejor su educación, sus resultados y mejorar su rendimiento.

Analizar los datos y la información de los estudiantes para clasificarlos, encontrar información que permita tomar mejores decisiones o mejorar el desempeño de los estudiantes es un campo de investigación interesante, que se enfoca principalmente en analizar y comprender los datos educativos de los estudiantes y generar reglas, clasificaciones y predicciones específicas para ayudar a los estudiantes en su desempeño educativo futuro.

La clasificación es la técnica de minería de datos más familiar y eficaz que se utiliza para clasificar y predecir valores. La minería de datos educativos (EDM) no es una excepción a este hecho, por lo tanto, se utilizará en este proyecto para analizar la información recopilada de los estudiantes a través de sus años de estudio y proporcionar clasificaciones basadas en los datos recopilados para predecir y clasificar a aquellos estudiantes que pueden continuar su proceso y/o transición educativa desde sus años de estudio en pregrado hacia los diferentes posgrados de la Pontificia Universidad Javeriana Cali. El objetivo del trabajo que se propone en este documento es generar un modelo de predicción que permita predecir la continuidad en el proceso de formación académica en posgrado de los estudiantes egresados de pregrado de la Pontificia Universidad Javeriana Cali, a partir de la identificación de atributos relacionados, a través de técnicas de minería de datos. Este conocimiento puede ayudar a las áreas de Promoción Estudiantil a mejorar sus estrategias de atracción, teniendo un mayor enfoque en aquellos estudiantes que presentan una alta certeza o grado de probabilidad de continuar su formación académica de posgrados en la universidad.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

Desde el año 2018, la cantidad de estudiantes matriculados en Posgrados en la Pontificia Universidad Javeriana Cali ha tenido una tendencia a la baja; en la Facultad de Ingeniería y Ciencias en los rangos de matriculados desde el segundo periodo del 2018 al segundo periodo del 2020 se ha tenido una caída de casi un 30% en el número de estudiantes. Este es un desafío permanente para las instituciones de educación superior porque la deserción estudiantil trae consigo problemas económicos y de reputación para las universidades. Frente a esta problemática, han sido desarrolladas una variedad de estrategias de retención para revertir esta creciente tendencia regional, nacional y global.

La pandemia de coronavirus que afecta al mundo desde el año 2020 obligó a muchas instituciones de educación superior a replantear sus metodologías de atracción hacia los nuevos integrantes en los diferentes niveles académicos de pregrado y posgrados. La transformación digital surgió entonces como una alternativa para que las instituciones pudieran reinventarse y seguir ofreciendo las diferentes alternativas de educación a los diferentes públicos. El crecimiento de la virtualidad supuso una alternativa para mantener la productividad, evitar la exposición al contagio, los traslados y gestionar el tiempo de forma eficiente. Sin embargo, también supuso una nueva forma de interacción en el mercado para la oferta de los programas académicos. Esta oferta significaba la explotación de los diferentes medios de comunicación y redes sociales para poder llegar a los clientes de forma oportuna y asertiva.

Según el informe “Big data en educación” desarrollado por la Universidad EAFIT y la Red de Inteligencia Competitiva, el país que más ha investigado el tema es Estados Unidos, mostrando que mucha de esta investigación se ha aplicado a casos reales en sus propias escuelas y universidades [1]. Igualmente, China y Reino Unido son dos países que ya han comenzado a estudiar el tema y a ver su importancia en el mejoramiento de la calidad y pertinencia educativa que le están proporcionando a sus estudiantes. Por otro lado, la aplicación de la minería de datos en la educación viene teniendo un importante lugar dentro del conjunto de investigaciones que buscan desarrollar métodos para explorar la información que se genera dentro de los ambientes educativos con el objetivo de entender la forma en que los estudiantes aprenden, para poder tomar las decisiones adecuadas que garanticen el éxito en el proceso educativo[2].

Por tanto, este trabajo de grado pretende realizar un estudio a partir de la aplicación de técnicas de minería de datos para generar un modelo de predicción a partir de la selección y evaluación de diferentes técnicas que permitan la interpretabilidad de dicho modelo, logrando predecir tendencias y patrones de comportamiento en los estudiantes de pregrado, para tener

continuidad académica en posgrados de la Pontificia Universidad Javeriana Cali. Además, hacer aprovechamiento de data mining para revelar las características de los estudiantes y predecir su tránsito de pregrado a posgrado que permita continuar con su formación en conocimientos específicos y especializados de su carrera profesional.

1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo aplicar técnicas de aprendizaje automático para generar un modelo de predicción de continuidad académica de formación en posgrados de estudiantes de pregrado, para aportar en el proceso de atracción y oferta académica de posgrados en el área de mercadeo de la Dirección de Promoción de la Pontificia Universidad Javeriana Cali?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Generar un modelo de predicción que permita predecir la continuidad en el proceso de formación académica en posgrado de los estudiantes egresados de pregrado de la Pontificia Universidad Javeriana Cali, a partir de la identificación de atributos relacionados, a través de técnicas de minería de datos.

2.2. OBJETIVOS ESPECÍFICOS

Identificar, recopilar, anonimizar, integrar y limpiar los datos obtenidos de los sistemas de información con el objetivo de establecer las relaciones que permitan definir la continuidad de los estudiantes egresados de pregrado.

Aplicar las técnicas de aprendizaje automático para predecir si un estudiante egresado de pregrado continuará sus estudios de posgrado en la Pontificia Universidad Javeriana Cali.

Evaluar y comparar los modelos generados para seleccionar el más apropiado para la predicción.

2.3. RESULTADOS OBTENIDOS

Como resultado de este trabajo, se obtuvo un modelo aplicando técnicas de aprendizaje automático, que permite predecir la continuidad en estudios de posgrado de estudiantes egresados de pregrado de la Facultad de Ingeniería y Ciencias, enmarcado en la estrategia de la unidad organizacional de las Oficinas de Promoción Institucional y Mercadeo de la Pontificia Universidad Javeriana Cali.

3. JUSTIFICACIÓN

Los métodos y técnicas de minería de datos se han convertido en un área de investigación importante en los últimos años, porque su inclusión en un proceso de análisis de datos puede revelar relaciones ocultas, patrones de comportamiento, perfiles de entidades y regularidades similares en los datos almacenados en grandes bases de datos.

El conocimiento descubierto por la inteligencia artificial difícilmente podría adquirirse por medios tradicionales, como el análisis estadístico, la consulta de datos u otros métodos analíticos, debido a la gran cantidad de datos recopilados y la vaga idea de la existencia del conocimiento. Por lo tanto, el descubrimiento de conocimiento en datos (KDD) y la minería de datos (DM) como parte integral, son indispensables para los analistas de datos. Sin embargo, lo mencionado anteriormente es cierto solo en el caso de datos de entrada de calidad, o datos que podrían transferirse a través del proceso de preprocesamiento[5]. En un artículo publicado por el Servicio de Investigación del Congreso de los Estados Unidos (CRS) declararon que: "La calidad de los datos es un tema multifacético que representa uno de los mayores desafíos para la minería de datos". Se ha reconocido que el éxito de todo un proceso KDD depende de las entradas proporcionadas[6].

En los últimos años, las instituciones de educación han incrementado cada vez más el uso de análisis de datos para investigar cuestiones científicas dentro de la investigación del dominio educativo para comprender mejor a los estudiantes y sus comportamientos de aprendizaje. Este análisis de datos va muy de la mano del crecimiento de herramientas tecnológicas que se combinan con el aprendizaje tradicional, a fin de generar estrategias para descubrir conexiones ocultas o previamente desconocidas y generar hipótesis en profundidad[8].

Una vez se logra consolidar la información a través del uso de herramientas de análisis de minería de datos, entra a jugar un papel importante el análisis predictivo; analizando datos demográficos y de rendimiento estudiantil, las instituciones de educación pueden predecir diferentes aspectos en el rendimiento de los estudiantes o pueden guiar sus estrategias en consolidar el mercadeo para obtener mejores resultados en la atracción de nuevos estudiantes y aumentar la tasa en la retención de sus estudiantes en las modalidades presenciales y en línea.

Según Gartner[4], “El análisis predictivo es una forma de análisis avanzado que examina los datos o el contenido para responder a la pregunta "¿Qué va a pasar?" o más precisamente, "¿Qué es probable que suceda?", y se caracteriza por técnicas como análisis de regresión, pronóstico, estadísticas multivariantes, coincidencia de patrones, modelado predictivo y pronóstico.”

El análisis predictivo utiliza técnicas de minería de datos, estadística, modelización, aprendizaje automático e inteligencia artificial, para analizar los datos actuales y hacer predicciones sobre el futuro. La técnica de minería de datos es usada con la finalidad de obtener información específica

que se encuentra oculta en grandes volúmenes de información y que aporta características y conocimientos útiles para las instituciones de educación.

Por otro lado, cuando la información no cuenta con una serie de datos ya etiquetados en diferentes categorías, grupos o clases en base a las cuales se puedan hacer predicciones o aplicar análisis predictivos, es necesario recurrir a técnicas de análisis de clúster o clustering.

Finalmente, los cambios en la educación superior y la movilidad estudiantil, hacen que la oferta educativa sea cada vez más competitiva, dinámica y se ajuste a las necesidades de los diferentes aspirantes de los posgrados. Por lo anterior, es importante proponer y desarrollar técnicas que permitan tomar decisiones oportunas frente al mercadeo educativo, aunque no se encontraron trabajos que precisen puntualmente sobre investigaciones previas acerca de la continuidad del proceso de formación de estudiantes de pregrado, es importante la aplicación de estas técnicas orientadas hacia un mayor grado de certeza en los resultados de las diferentes ofertas de atracción y marketing de la universidad; si bien es cierto que lo que más puede primar en la decisión de un estudiante para continuar realizando posgrados en la universidad es un plan de estudio bien estructurado, que permita satisfacer los objetivos de su vida profesional, pueden existir mecánismos o estrategías, como los modelos de predicción, que generen valor agregado a la institución y ser claves en el proceso de atracción de dichos estudiantes.

4. MARCO TEÓRICO DE REFERENCIA Y ANTECEDENTES

4.1. MINERÍA DE DATOS

La minería de datos (DM) es una técnica dedicada a escanear grandes conjuntos de datos, generar información y descubrir conocimiento a partir de esos datos. El significado del término minería tradicional sesga los motivos de la minería de datos. Pero, en lugar de buscar minerales naturales, el objetivo es el conocimiento. DM busca descubrir patrones de datos, organizar información de relaciones ocultas, estructurar reglas de asociación, estimar valores de elementos desconocidos para clasificar objetos, componer grupos de objetos homogéneos y revelar muchos tipos de hallazgos que no son fácilmente producidos por un sistema de información clásico. Por lo tanto, los resultados de la DM representan un apoyo valioso para la toma de decisiones [2].

La educación es un nuevo objetivo de aplicación de DM para el descubrimiento de conocimientos, la toma de decisiones y la recomendación. Hoy en día, el uso de DM en el ámbito educativo es incipiente y da origen al campo de investigación de la minería de datos educativos (EDM)[10].

La EDM surge como un paradigma orientado a diseñar modelos, tareas, métodos y algoritmos para explorar datos de contextos educativos. EDM busca descubrir patrones y hacer predicciones que caracterizan los comportamientos y logros de los estudiantes, el contenido del conocimiento del dominio, las evaluaciones, las funcionalidades educativas y las aplicaciones.

4.2. APRENDIZAJE AUTOMÁTICO

Machine Learning es un área de la inteligencia artificial que engloba un conjunto de técnicas que hacen posible el aprendizaje automático a través del entrenamiento a partir de conjuntos de datos. Hoy en día existen diferentes modelos que utilizan esta técnica y consiguen una precisión incluso superior a la de los humanos en las mismas tareas, por ejemplo en el reconocimiento de objetos en una imagen [11].

La construcción de modelos de aprendizaje automático requiere adaptaciones propias debido a la naturaleza de los datos o a la problemática a la que se aplica. Así, surge la necesidad de investigar las diferentes técnicas que permitan obtener resultados precisos y confiables en un tiempo razonable.

El aprendizaje automático permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita [12]. Sin embargo, no es un proceso sencillo. Conforme el algoritmo ingiere datos de entrenamiento, es posible producir modelos más precisos basados en

datos. Un modelo de aprendizaje automático es la salida de información que se genera cuando se entrena un algoritmo de aprendizaje automático con datos. Después del entrenamiento, al proporcionar un modelo con una entrada, se le dará una salida. Por ejemplo, un algoritmo predictivo creará un modelo predictivo. A continuación, cuando proporcione el modelo predictivo con datos, recibirá un pronóstico basado en los datos que entrenaron al modelo.

El aprendizaje automático permite entrenar modelos de manera iterativa con conjuntos de datos antes de ser implementados. Este proceso iterativo de modelos conduce a una mejora en los tipos de asociaciones hechas entre los elementos de datos. Debido a su complejidad y tamaño, estos patrones y asociaciones podrían haber sido fácilmente pasados por alto por la observación humana. Después de que un modelo ha sido entrenado, se puede utilizar en tiempo real para aprender de los datos. Las mejoras en la precisión son el resultado del proceso de entrenamiento y la automatización que forman parte del Machine Learning [12].

4.2.1 Tipos de aprendizaje automático. Las técnicas de aprendizaje automático son muy utilizadas hoy en día para mejorar la precisión de los modelos predictivos. Dependiendo de la naturaleza del problema empresarial que se está atendiendo, existen diferentes enfoques basados en el tipo y volumen de los datos.

- **Aprendizaje supervisado.** El aprendizaje supervisado comienza típicamente con un conjunto establecido de datos y una cierta comprensión de cómo se clasifican estos datos. El aprendizaje supervisado tiene la intención de encontrar patrones en datos que se pueden aplicar a un proceso de analítica. Estos datos tienen características etiquetadas que definen el significado de los datos. Por ejemplo, se puede crear una aplicación de aprendizaje automático con base en imágenes y descripciones escritas que distinga entre varias especies de animales.

- **Aprendizaje no supervisado.** El aprendizaje no supervisado se utiliza cuando el problema requiere una cantidad masiva de datos sin etiquetar. Por ejemplo, las aplicaciones de redes sociales, tales como Twitter, Instagram y Snapchat, tienen grandes cantidades de datos sin etiquetar. La comprensión del significado detrás de estos datos requiere algoritmos que clasifican los datos con base en los patrones o clústeres que encuentran. El aprendizaje no supervisado lleva a cabo un proceso iterativo, analizando los datos sin intervención humana. Se utiliza en la detección de spam en e-mails. Existen demasiadas variables en los emails legítimos y de spam para que un analista etiquete una cantidad masiva de e-mail no solicitado. En su lugar, los clasificadores de Machine Learning, basados en Clustering y asociación, se aplican para identificar e-mail no deseado.

- **Aprendizaje de refuerzo.** El aprendizaje de refuerzo es un modelo de aprendizaje conductual. El algoritmo recibe retroalimentación del análisis de datos, conduciendo al usuario hacia el mejor resultado. El aprendizaje de refuerzo difiere de otros tipos de aprendizaje supervisado, porque el sistema no está entrenado con el conjunto de datos de ejemplo. Más bien, el sistema aprende a través de la prueba y el error. Por lo tanto, una secuencia de decisiones exitosas conduce al fortalecimiento del proceso, porque es el que resuelve el problema de manera más efectiva.

- **Aprendizaje profundo.** El aprendizaje profundo es un método específico de Machine Learning que incorpora las redes neuronales en capas sucesivas para aprender de los datos de manera iterativa. El Deep Learning es especialmente útil cuando se trata de aprender patrones de datos no estructurados. Las redes neuronales complejas de aprendizaje profundo están diseñadas para emular cómo funciona el cerebro humano, así que las computadoras pueden ser entrenadas para lidiar con abstracciones y problemas mal definidos. Las redes neuronales y el aprendizaje profundo se utilizan a menudo en el reconocimiento de imágenes, voz y aplicaciones de visión de computadora.

4.2.2 Validación cruzada: elección de parámetros en un algoritmo. Uno de los pasos importantes a tener en cuenta cuando se desea entrenar un algoritmo de aprendizaje automático supervisado es estimar cuál es su comportamiento, es decir, con qué precisión clasificará el modelo creado con un algoritmo elegido aquellos datos nuevos que no han sido vistos previamente. Un conocido error metodológico para estos casos es el de sobreentrenar el algoritmo con unos datos para los que ya se conoce el resultado deseado, esto es comúnmente conocido como *overfitting*. Por otro lado, es posible que el modelo sea demasiado simple y no se cuente con los valores adecuados en su entrenamiento, esto es conocido como *underfitting* [12]. Por lo anterior, es necesario encontrar la mejor forma de evaluar el modelo cuidadosamente y, de esta forma, asegurar que se va a obtener el mejor rendimiento del mismo. Para ello existen técnicas de validación cruzada, comúnmente conocidas como *cross-validation*, que indican cómo de preciso se comportará el algoritmo para datos nuevos no observados.

La validación cruzada es una técnica que se usa para poder evaluar los resultados de un análisis para que así se pueda garantizar que son independientes de la partición entre los datos que se usan del conjunto de entrenamiento y los datos del conjunto de prueba [12]. Esta técnica consiste en evaluar el modelo mediante el entrenamiento de varios modelos en subconjuntos de los datos de entrada disponibles y evaluarlos con el subconjunto complementario de los datos. Generalmente se utiliza la validación cruzada para detectar el sobreajuste, es decir, en aquellos casos en los que no se logre generalizar un patrón.

4.2.3 Métricas de evaluación del rendimiento de un modelo. Estas métricas permiten elegir cuál es el mejor de todos los algoritmos evaluando los distintos valores obtenidos de las métricas [12]. Para entender las distintas métricas es necesario entender previamente de donde se extraen las mismas. Para ello es necesario entender con anterioridad el concepto de matriz de confusión: una matriz que ayuda a conocer cuál ha sido el rendimiento de un algoritmo. Esta matriz muestra en un cuadro los siguientes valores:

- **Verdaderos Positivos (VP).** Es la cantidad de aquellas observaciones que fueron clasificados como pertenecientes a una clase y el modelo acertó.
- **Falsos Positivos (FP).** Esas observaciones que no pertenecían a una clase pero se llegaron a considerar perteneciente a ellas.
- **Falsos Negativos (FN).** Aquellas observaciones que pertenecían a una clase pero no se consideraron en ella.
- **Verdaderos Negativos (VN).** Es la cantidad de aquellas observaciones que no pertenecían a una clase y el modelo las clasificó correctamente.

A partir de estos 4 valores previos se pueden sacar las métricas de rendimiento que sirven para medir el rendimiento de un modelo:

- **Exactitud**. La exactitud de la clasificación es la relación entre las predicciones correctas y el número total de predicciones. O más simplemente, con qué frecuencia es correcto el clasificador.

- **Precisión**. La precisión es la relación entre las predicciones correctas y el número total de predicciones correctas previstas. Esto mide la precisión del clasificador a la hora de predecir casos positivos.

- **Sensibilidad**. La sensibilidad también es llamada *recall*, es la relación entre las predicciones positivas correctas y el número total de predicciones positivas. O más simplemente, cuán sensible es el clasificador para detectar instancias positivas. Esto también se conoce como la tasa verdadera positiva.

- **Puntaje F1**. El puntaje F1 es el promedio ponderado de precisión y sensibilidad. Por lo tanto, esta puntuación tiene en cuenta tanto los falsos positivos como los falsos negativos.

4.3. TÉCNICAS DE CIENCIA DE DATOS O ANÁLISIS DE DATOS

Las técnicas de Data Science o Data Analytics que tanto interés despiertan hoy en día, en realidad surgieron en la década de los 90, cuando se usaba el término KDD (Knowledge Discovery in

Databases) para referirse al (amplio) concepto de hallar conocimiento en los datos. En un intento de normalización de este proceso de descubrimiento de conocimiento, de forma similar a como se hace en ingeniería software para normalizar el proceso de desarrollo software, surgieron a finales de los años 90 dos metodologías principales: CRISP-DM (Cross Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model, and Assess). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase[13].

En 2015, IBM Corporation, uno de los impulsores tradicionales de CRISP-DM, planteó una nueva metodología llamada Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM) que extiende CRISP-DM, y es parte de la metodología general ASUM (Analytics Solutions Unified Method) incorporada en los productos y soluciones analíticas de IBM[13].

4.4. CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP-DM establece un proyecto de minería de datos a través de un modelo de proceso iterativo, desarrollado en 1996, y actualmente es la metodología más favorecida desde entonces[14]. CRISP-DM comprende las siguientes fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue (Fig. 1). Generalmente, estas tareas se establecen como fases secuenciales, pero dentro de esta corriente se presentan muchos ciclos iterativos.

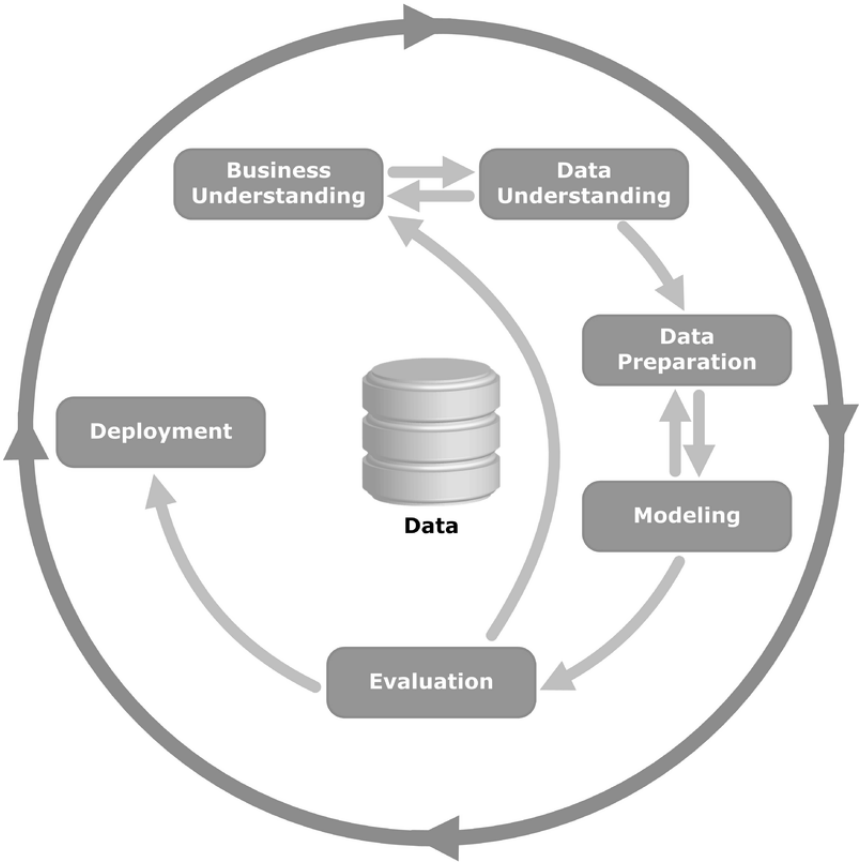


Fig. 1 Ciclo de vida del proceso de CRISP-DM. [15]

4.4.1 FASE I: ENTENDIMIENTO DEL NEGOCIO. Esta primera fase es probablemente la más importante y agrupa las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de la minería de datos, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados.

4.4.2 FASE II: ENTENDIMIENTO DE LOS DATOS. Esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las dos siguientes fases son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos específica para el proyecto de DM (Data Mining), ya que durante el desarrollo del proyecto es posible que se generen frecuentes y abundantes accesos a la base de datos con el fin de realizar consultas y

probablemente se produzcan modificaciones, lo cual podría generar muchos problemas.

4.4.3 FASE III: PREPARACIÓN DE LOS DATOS. En esta fase y una vez efectuada la recolección inicial de los datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente, estas pueden ser técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para explotación de los datos. La preparación de los datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Esta fase se encuentra relacionada con la fase de modelado, ya que, en función de la técnica de modelado elegida, los datos requieren ser procesados de una manera o de otra, por esta razón las fases de preparación y de modelado interactúan de forma permanente.

4.4.4 FASE IV: MODELADO. En esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas para utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada para el problema.
- Disponer de los datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Previamente al modelado de los datos se debe determinar un método de evaluación de los modelos que permita establecer el grado de adecuación de cada uno de ellos. Después de concluir estas tareas genéricas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo.

4.4.5 FASE V: EVALUACIÓN. En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se pueda haber cometido algún error. Hay que considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo.

4.4.6 FASE VI: DESPLIEGUE. En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el

analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso, como, por ejemplo, en análisis de riesgo crediticio, detección de fraudes, etc. Generalmente un proyecto de Data Mining no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento. Por otra parte, en esta fase se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados. Las tareas involucradas son las siguientes: plan de implementación (esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación), monitorización/mantenimiento (si los modelos resultantes del proceso de Data Mining son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos), producción del informe final (dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto) y la revisión del proyecto (en este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar).

4.5. BIG DATA

Big Data es el tipo de datos que tienen una escala muy grande, contienen una mayor variedad, se presentan a una velocidad superior y requieren un tipo especial de almacenamiento y técnicas para almacenarlos. Adicionalmente, requiere algoritmos y técnicas especiales para procesarlos y obtener información útil a partir de datos sin procesar. Como este concepto cobró un gran impulso en los últimos años, tiene muchos desafíos, oportunidades y hay muchas tecnologías emergentes para el análisis de Big Data. Dado que el Big Data es de gran escala y los datos provienen de diversas fuentes y con diferentes formatos (Fig. 2), los sistemas de gestión de datos convencionales son incapaces de procesar Big Data[9].

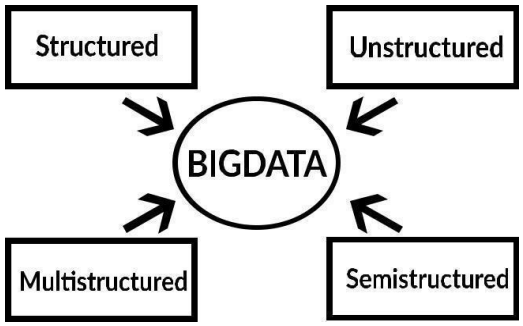


Fig. 2 Tipos de datos en Big Data [9]

4.5.1 Datos estructurados. El tipo de datos en forma de tablas y filas con los nombres de atributos adecuados y estos datos se pueden mostrar fácilmente en hojas de cálculo.

4.5.2 Datos sin estructura. El tipo de datos que no son tablas y filas con nombres de atributos adecuados y es realmente difícil mostrar datos en una hoja de cálculo.

4.5.3 Datos multi estructurados. El tipo de datos que tienen diferente tipo de estructura significa que tienen algunos datos numéricos, algunos datos en forma de imágenes, etc.

4.5.4 Datos semiestructurados. El tipo de datos que tienen algunos datos estructurados y algunos datos sin la estructura adecuada puede deberse a una mezcla de datos de base de datos y archivos de registro. Los datos de todas estas categorías se almacenan juntos en Big Data para análisis, por lo que es realmente difícil almacenar estos datos juntos.

4.5.5 Dimensiones. Big Data tiene 3 dimensiones principales por sus características que se pueden describir de la siguiente manera: volumen, variedad y velocidad (Fig. 3). Volumen hace referencia a la cantidad de datos que llegan para análisis y es un desafío realmente difícil mantenerlos, almacenarlos en algún lugar y analizarlos. La variedad en Big Data significa que se tienen diferentes tipos de datos, puede tener datos de video, puede tener datos textuales, datos binarios, audio, imágenes, numéricos, etc. En Big Data, los datos tienen diferentes variedades, depende de las fuentes de dónde vienen los datos. Y finalmente velocidad se refiere al ritmo en que los datos de entrada fluyen desde las diversas fuentes como procesos de negocio, máquinas y sensores, redes sociales, dispositivos móviles, etc.

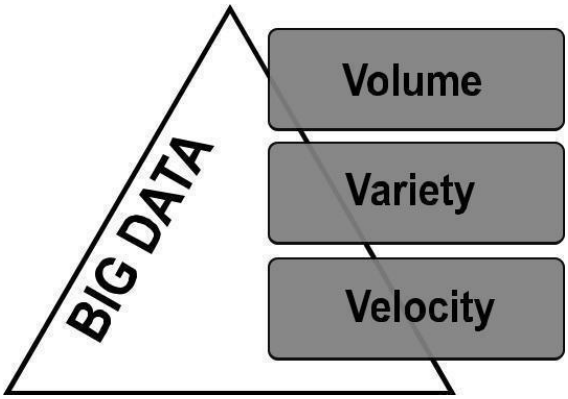


Fig. 3 Dimensiones en Big Data [9]

4.6. TRABAJOS RELACIONADOS

La Universidad Estatal de Arizona (la más innovadora de Estados Unidos, según los últimos tres rankings publicados por U.S. News & World Report), implementa un sistema muy particular para monitorear a sus estudiantes[1]. La lógica es bastante simple, pero tiene implicaciones enormes: cada semestre, los alumnos "primíparos" reciben un carnet institucional que les sirve para todo. Con él ingresan a los edificios de clase, a la biblioteca, a las residencias y al centro recreativo; compran, como con una tarjeta de crédito, víveres en las tiendas del campus y entradas para el teatro más cercano, y hasta lo usan en las máquinas de gaseosas y dulces de la universidad. A cambio, la institución recibe información constante de dónde están, a qué hora y con quién.

"Es como un sensor que llevan puesto y que podemos usar para seguirles la pista. Aunque los carnets no están hechos para monitorear sus interacciones sociales, cruzando el tiempo exacto que están dos estudiantes en una ubicación puedes hacerlo", dijo en un comunicado oficial Sudha Ram, directora del Centro de Inteligencia y Analítica de Negocios (Insite, por sus siglas en inglés) de la universidad.

Ram ideó este programa de análisis de Big Data y empezó a aplicarlo hace tres años. Su objetivo es disminuir el factor de riesgo de deserción, pues, como encontró la directora, la integración social y la rutina de estudios de los estudiantes predicen mejor que las notas cuando un alumno de primer año va a desertar. Y, hasta ahora, lo ha logrado. "Tras solo 12 semanas de clase, somos capaces de identificar previamente el 85% de los estudiantes que desertarán a final del primer semestre", señala Ram. En parte gracias a este dato, y a una intervención focalizada en los alumnos bajo riesgo, en el año 2017 la institución rompió su récord de retención estudiantil con un 85,6%, superando el promedio nacional (76%). Y, aunque este no es solo un logro del proyecto del Insite, sí demuestra lo que puede llegar a lograr el uso de learning analytics. Desde el punto de vista técnico es interesante, pero desde la privacidad es cuestionable.

Middle East College (MEC) es una de las instituciones de educación superior privadas más grandes de Omán. Student Success Center (SSC) es un departamento de apoyo en MEC para monitorear el progreso de los estudiantes[1]. El objetivo principal de este departamento es identificar los factores que afectan el desempeño de los estudiantes académicamente débiles y brindar intervenciones adecuadas en asociación con otros departamentos. MEC está afiliado a la Universidad de Coventry, Reino Unido y la Universidad de Wolverhampton, Reino Unido. Los módulos en varios programas de pregrado se clasifican en diferentes niveles como Nivel 1, Nivel 2 y Nivel 3 para incorporar habilidades de pensamiento de orden inferior y superior de acuerdo con la taxonomía de Bloom.

El estudio se realizó para determinar la asociación de la edad y el género con el desempeño en matemáticas entre estudiantes de secundaria en EE. UU. El estudio reveló que los puntajes de GPA en matemáticas disminuyeron con la edad. El trabajo de investigación utilizó técnicas de minería de datos para predecir los resultados de los estudiantes nuevos mediante el análisis del género, las notas obtenidas en el examen de calificación anterior y los exámenes de ingreso. El análisis utilizó atributos como género, raza, ciudad natal, ingreso familiar y modo de ingreso a la universidad para determinar el desempeño del estudiante utilizando varias técnicas de minería de datos. Esta investigación notó que la raza era el factor más importante que determinaba el desempeño del estudiante, seguido por el ingreso familiar.

AltSchool, una start up educativa de Estados Unidos ha puesto en marcha un ambicioso proyecto de Big Data con el fin de mejorar la educación de los estudiantes de 0 a 12 años[16].

El grupo escolar cuenta para ello con aplicaciones que controlan la asistencia de los alumnos y ordenadores y otras herramientas tecnológicas que registran permanentemente su actividad académica. Además, disponen de cámaras de vídeo para grabar constantemente lo que sucede en las aulas desde múltiples ángulos, con el fin de capturar las expresiones faciales de los niños, registrar su forma de hablar, el vocabulario, qué cosas les llaman más la atención, etc. El análisis de toda esa información proporciona una comprensión integral de cada alumno basada en sus patrones de conducta, estados de ánimo, rendimiento, etc., que permite darle a cada uno la educación que necesita atendiendo a sus necesidades y diferencias.

En Colombia, las pruebas de Estado Saber-Pro han sido diseñadas para apoyar la evaluación y el mejoramiento de la educación superior en el país. Aplicando metodologías de minería de datos, se realizó un estudio de los resultados obtenidos en las pruebas Saber-Pro[17] de estudiantes de la Facultad de Ingeniería en Antioquia (Colombia). Este estudio obtuvo como resultado que se encuentra que algunas de las variables más influyentes sobre el resultado de las pruebas son: el número de personas a cargo, método de enseñanza, si el hogar es permanente, el carácter académico de la institución y facilidades económicas como tener horno micro gas y motocicleta.

5. APLICACIÓN DE LA METODOLOGÍA CRISP-DM

Con el fin de dar cumplimiento al objetivo de poder predecir la continuidad en el proceso académico de un estudiante egresado de pregrado bajo un conjunto de características relevantes de los estudiantes en su proceso de formación en la Pontificia Universidad Javeriana Cali, se aplicó la metodología CRISP-DM para el desarrollo de este proyecto siguiendo las fases que define la metodología y aplicando algunas actividades particulares de cada una de las fases.

5.1. FASE I: ENTENDIMIENTO DEL NEGOCIO

En esta primera fase se desarrollaron algunas actividades que permitieron conocer el contexto del proceso de oferta, atracción y promoción de la Pontificia Universidad Javeriana Cali:

- Reunión inicial con la directora de Promoción Institucional para determinar los objetivos estratégicos en la atracción de aspirantes de programas de posgrados donde se tuvo un primer acercamiento al proceso y a los atributos relevantes para tener en cuenta en los egresados. El resultado de la reunión fue conocer las diferentes estrategias (pautas, envíos masivos de correos electrónicos, planes de referidos, etc.) que tiene el área para todo el tema de atracción de los posgrados. Se conocieron los diferentes perfiles de los programas de posgrados, así como aquellos programas que tienen una alta proporción de egresados de la universidad. Finalmente se conocieron algunos atributos que se tienen en cuenta al momento de lanzar las diferentes campañas de los programas para poder llegar a un público objetivo amplio a partir de la aplicación de las estrategias de mercadeo.
- Reunión con la directora de Promoción Institucional, el coordinador de Análisis de Datos y la coordinadora de Mercadeo para conocer con más profundidad las estrategias de engagement (estrategia utilizada en marketing digital para crear relaciones sólidas y duraderas con los usuarios) a partir de la atomización -rango de edades, intereses, profesiones, regiones y ciudades- de los diferentes públicos objetivos. Se conoció que la universidad le está apostando en los 2 últimos años al marketing digital, orientando y enfocando sus estrategias en las pautas digitales y a las diferentes herramientas de lookalike (su finalidad es encontrar usuarios similares o que presenten algún tipo de afinidad con aquellos que componen el target de una campaña para así, poder plantear un acercamiento a ellos y obtener un mayor alcance).
- Evaluación de la situación actual a partir de la información recolectada de la reunión inicial de contextualización de la operación de mercadeo y promoción.

A partir de las actividades mencionadas anteriormente se estableció el plan de proyecto, recolectando la información necesaria y alineada hacia el objetivo de predecir si un estudiante egresado continuará sus estudios de posgrado. La definición anterior surgió a partir de la revisión

en conjunto de los cierres de campaña de posgrados en el periodo 2021-2, donde se evidenció una gran oportunidad de direccionar campañas de promoción y contar con planes de mercadeo para atraer con una mayor fuerza a los egresados de la universidad, con el objetivo de lograr tener una consolidación temprana de prospectos que pueden llegar a estudiar posgrados.

5.2. FASE II: ENTENDIMIENTO DE LOS DATOS

En esta segunda fase se realizó la recolección inicial de los datos de los estudiantes que sirvieron como base para el desarrollo de la minería de datos, se determinó la cantidad de datos involucrados, el tipo de datos, su descripción y se establecieron las primeras relaciones más relevantes acerca de su edad, factor económico, programa académico, planes de estudio, etc.

5.2.1 FUENTES DE LOS DATOS. En esta etapa se trabajó con diferentes fuentes con información específica de los estudiantes. Se obtuvieron los datos de los sistemas institucionales de información académica y financiera del estudiante, así como de información que reposa en las estructuras de datos del proyecto institucional de BI. A continuación, se brinda una breve descripción de las fuentes de los datos que se pueden obtener de los diferentes sistemas mencionados anteriormente:

- **PeopleSoft Campus Solutions (CSS)**. Campus Solutions es parte de PeopleSoft ERP Suite y es la aplicación institucional de la Pontificia Universidad Javeriana Cali para almacenar y mantener los datos de los estudiantes. La aplicación contiene diferentes módulos como Admisiones, Comunidad del Campus, Registros Estudiantiles, Ayudas Financieras y Finanzas Estudiantiles. De esta fuente de datos se logró extraer información académica y financiera de los egresados (aquellos estudiantes que se graduaron en un periodo comprendido entre el 2006 y el 2020).

- **PeopleSoft Human Capital Management (HCM)**. Human Capital Management también es parte de PeopleSoft ERP Suite. A diferencia de CSS, HCM es utilizada para almacenar y mantener los datos de los empleados. La aplicación contiene diferentes módulos como Administración de Personal, en donde se configura y administra la contratación de los estudiantes que han sido monitores.

- **Business Intelligence (BI)**. El Proyecto de Business Intelligence y Data Mining de la Pontificia Universidad Javeriana Cali permite consolidar, transformar y procesar los datos a través de estrategias y herramientas que sirven para transformar información en conocimiento, con el objetivo de mejorar el proceso de toma de decisiones a nivel académico y administrativo. El proyecto recopila todo tipo de información de los estudiantes como sus datos demográficos, notas, asignaturas, etc., apoyando las estrategias de promoción institucional de la universidad.

5.2.2 DEFINICIÓN DEL CONJUNTO DE DATOS. En esta etapa se extrae la información relevante de los estudiantes (producto del análisis posterior a las reuniones sostenidas con las áreas de planeación, mercadeo y gestión de datos de la universidad) de las fuentes de datos mencionadas anteriormente. El periodo de tiempo del cual se hizo la extracción de la información fue desde el 01 de enero de 2006 hasta el 31 de diciembre de 2020, obteniendo 11.896 registros que fueron exportados inicialmente a un archivo de formato de libro de Excel (XLSX) para filtrar un total de 3620 registros pertenecientes a estudiantes egresados de la facultad de Ingeniería y Ciencias. Las variables asociadas al proyecto de minería de datos se muestran en la Tabla 1. Estas variables fueron las tenidas en cuenta inicialmente.

Tabla 1- Definición del Conjunto de Datos

Atributo	**Descripción**
Fecha de grado	Fecha de grado del estudiante. Este dato se utilizó para obtener la edad del estudiante al momento de graduarse de pregrado
Promedio académico	Promedio ponderado de las calificaciones obtenidas por el estudiante según el número de créditos de sus asignaturas cursadas
Ciudad	Ciudad de residencia del estudiante al momento de graduarse o último dato registrado
Género	Género del estudiante
Fecha de nacimiento	Fecha de nacimiento del estudiante
Lugar de nacimiento	Ciudad o municipio donde nació el estudiante
Tipo de identificación	Tipo de identificación del estudiante
Programa académico	Programa académico del estudiante que cursó y aprobó el plan de estudios
Movilidad	Atributo del estudiante que indica si cursó o no asignaturas de su plan de estudios en otras universidades de la ciudad o del país
Consejería académica	Atributo del estudiante que indica si recibió o no acompañamiento educativo por parte de docentes para su formación académica

Distinciones	Atributo del estudiante que indica si recibió o no distinciones académicas por los méritos extraordinarios relacionados a su progreso en la universidad
Créditos académicos	Cantidad de créditos matriculados del estudiante en su formación académica
Monitorias	Atributo del estudiante que indica si participó o no del programa de monitorias en la universidad
Valor de la matrícula	Valor total de pagos del estudiante por concepto de matrícula académica durante su ciclo de formación. Este dato se utilizó para obtener la proporción del valor de becas/descuentos y créditos financieros del estudiante
Becas/Descuentos	Valor total de becas y descuentos aplicados en la cuenta del estudiante para el incentivo en la continuidad de su proceso de formación académica
Beca continua	Atributo del estudiante que indica si recibió o no beca de manera ininterrumpida durante su ciclo de formación académica
Créditos financieros	Valor total de créditos monetarios que financian el valor de la matrícula del estudiante
Doble programa	Atributo del estudiante que indica si matriculó o no en algún momento de su ciclo de formación 2 programas académicos en forma simultánea del mismo nivel de formación de pregrado
Pruebas académicas	Atributo del estudiante que indica la cantidad de pruebas académicas en las que cayó. Un estudiante se encuentra en prueba si su promedio académico ponderado acumulado es inferior al establecido en el currículo o si el promedio ponderado del periodo es inferior a 2.5
Semestres matriculados	Número de semestres matriculados por el estudiante en el programa en el cual egresó. En la mayoría de los casos corresponde al número de semestres de duración de la carrera. Se tiene en cuenta los semestres en los cuales el estudiante matriculó asignaturas del programa de egreso sin tener en cuenta traslados de carreras.
Materias perdidas	Número de materias que perdió (reprobó) el estudiante durante su carrera
Estudios posgrado	Atributo del estudiante que indica si cursó o no algún programa de posgrado en la universidad

5.2.3 DESCRIPCIÓN DE LOS DATOS. En esta etapa se exploraron los datos seleccionados para

determinar los diferentes tipos de datos, frecuencias y medidas de centralización. Se conocieron los datos con los cuales se trabajó antes de empezar a realizar el preprocesamiento de los datos, conociendo las características del conjunto de datos obtenido en las etapas previas, creando tablas de frecuencia y gráficos de distribución de los datos.

En la Tabla 2 se muestra la información del set de datos con el que se trabajó indicando la cantidad de registros que contiene el set, los diferentes tipos de datos encontrados y la información acerca de la cantidad de valores faltantes por cada uno de los atributos.

Tabla 2 - Información del set de datos

Atributo	Cantidad de datos no faltantes	Tipo de dato
Fecha Grado	3620 non-null	datetime64[ns]
Promedio Académico	3620 non-null	float64
Ciudad	1249 non-null	object
Género	3620 non-null	object
Fecha Nacimiento	3619 non-null	datetime64[ns]
Lugar Nacimiento	3610 non-null	object
Tipo Identificación	3620 non-null	object
Programa Académico	3620 non-null	object
Movilidad	3620 non-null	object
Consejerías	3620 non-null	object
Distinciones	3620 non-null	object
Créditos Académicos	3620 non-null	int64
Monitorías	3620 non-null	object
Valor Matrícula	3616 non-null	float64
Becas/Descuentos	345 non-null	float64
Beca Permanente	3620 non-null	object
Créditos Financieros	634 non-null	float64
Doble Programa	3620 non-null	object
Pruebas Académicas	3620 non-null	int64
Semestres Matriculados	3620 non-null	int64
Materias Perdidas	3620 non-null	int64
Matrícula Posgrados	3620 non-null	object

En la Figura 4 se muestra la distribución de frecuencias de la clase mayoritaria y minoritaria en el conjunto de datos. Se presenta un desbalance pronunciado entre los datos, en donde la clase mayoritaria (3068 registros) pertenece a aquellos estudiantes que no continúan sus estudios de

posgrados en la universidad, mientras que la clase minoritaria (552 registros) corresponde a aquellos egresados que continúan sus estudios de posgrado en la universidad.

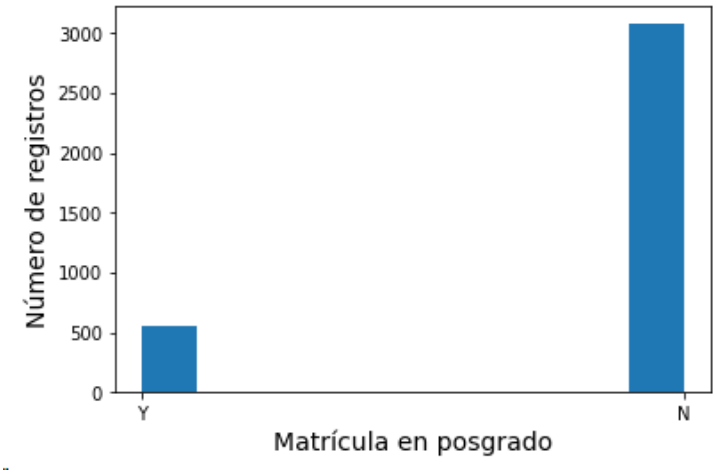


Fig. 4 Distribución de frecuencias de clases

En la Figura 5 se muestra la distribución de frecuencias por género en el conjunto de datos. Dentro del número de egresados, aproximadamente un 65% corresponde a egresados de género masculino, mientras que un 35% corresponde a egresados de género femenino. En la Figura 6 se muestra la distribución de frecuencias por programa académico en donde se observa que la predominancia de la carrera se da por parte de ingeniería industrial.

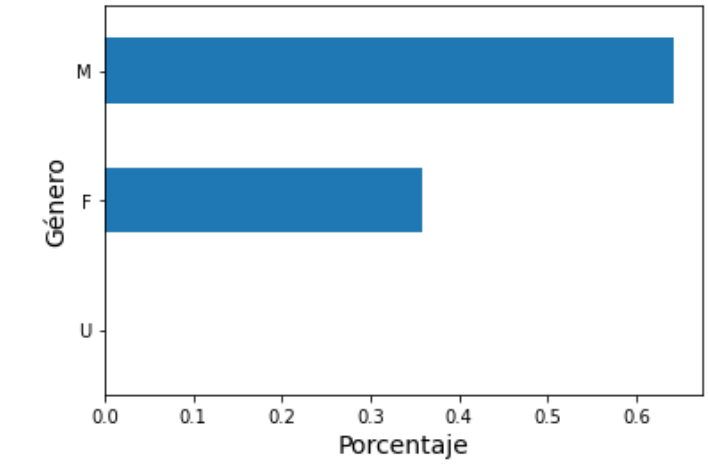


Fig. 5 Distribución de frecuencias por género

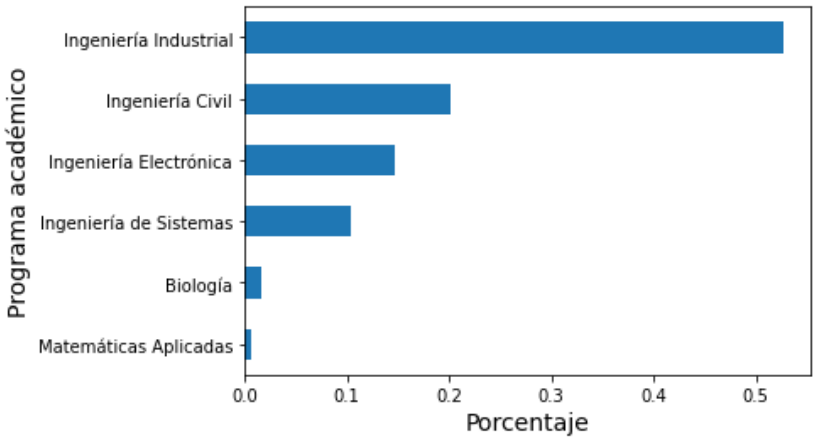


Fig. 6 Distribución de frecuencias por programa académico

En la Tabla 3 se muestran las medidas de centralidad, resumiendo la tendencia central, la dispersión y la forma de la distribución del conjunto de datos.

Al hacer el análisis inicial de los datos con los cuales se contaba, se tomó la decisión de realizar una depuración previa que consistió en la eliminación de registros muy atípicos y que pudieran llegar a generar ruido al momento del entrenamiento y evaluación de los modelos, ya que aunque el rendimiento de los modelos sea mejor para dicho conjunto de datos en particular, no sería aplicable para otros conjuntos de datos, teniendo en cuenta que los datos depurados inicialmente tienen una alta probabilidad a ser datos erróneos al no ajustarse a lo definido en el Reglamento del Estudiante.

Algunos de los datos eliminados corresponden a registros con promedios académicos menores a 3.25 (según el Reglamento del Estudiante un estudiante no se puede graduar si se encuentra en prueba académica), créditos académicos por debajo del 25% de los créditos correspondientes a cada uno de los programas académicos (una de las condiciones para graduarse un estudiante es haber cursado y aprobado en la universidad al menos el 25% de los créditos correspondientes al programa según el Reglamento del Estudiante) y cantidad de pruebas académicas que superan el 60% del total de semestres matriculados del estudiante (se tomó como límite dicho porcentaje teniendo en cuenta que algunos datos pudieran ser erróneos y que un estudiante de un programa académico, regularmente de 10 periodos académicos, es expulsado del programa si cae en prueba académica durante 3 periodos consecutivos según el Reglamento del Estudiante). Se tomó la decisión de dejar los datos recolectados de la fuente de la característica del número de materias perdidas durante la carrera, que aunque en algunos casos se registran valores elevados, puede ser una variación acorde al desempeño académico del estudiante y no se consideró como un valor erróneo o atípico.

Por otro lado, al hacer el análisis de los registros correspondientes al valor de la matrícula se encontraron valores muy pequeños, sin embargo, se tomó la decisión de dejarlos teniendo en

cuenta que no se presentó un impacto porcentual en las características de las becas y créditos financieros porque correspondían a estudiantes con datos faltantes en dichas características (el manejo de estos datos faltantes se presenta más adelante). En el análisis se tomó como referencia el valor promedio que puede pagar un estudiante durante su etapa de formación académica en pregrado y cuyos registros estuvieran por debajo del 25% de dicho valor, teniendo en cuenta lo mencionado en el párrafo anterior en relación con el porcentaje de créditos matriculados en la universidad necesarios para la graduación de un estudiante.

En la Tabla 3 se encuentran marcados con un asterisco aquellos atributos sobre los cuales se realizó el análisis y se tomó la decisión de eliminar registros.

Tabla 3 - Estadísticas descriptivas del conjunto de datos

	*Promedio académico	*Créditos académicos	Valor matrícula ($ COP)	Becas Descuentos ($ COP)	Créditos financieros ($ COP)	*Pruebas académicas	Semestres matriculados	Materias perdidas
count	3620.00	3620.00	3616.00	345.00	634.00	3620.00	3620.00	3620.00
mean	3.83	172.28	47993745.15	-10406848.41	-18775701	0.43	12.29	5.74
std	0.25	20.82	16579687.34	12745538.17	17114265	1.14	2.98	6.32
min	3.25	42.00	406000.00	-70944946.00	-75014809	0.00	2.00	0.00
25%	3.65	170.00	35014411.25	-14118418.00	-29040634	0.00	10.00	1.00
50%	3.79	177.00	47390074.50	-5598325.00	-13027331	0.00	12.00	4.00
75%	3.98	180.00	60554647.00	-1791133.00	-5168202	0.00	14.00	9.00
max	4.82	245.00	118985398	0.00	0.00	14.00	29.00	44.00

5.2.4 EXPLORACIÓN DE LOS DATOS. Se identificó que en las fases posteriores podría ser necesario aplicar la eliminación de la variable del tipo de identificación pues no contiene información de importancia debido a que no aporta valor al análisis teniendo en cuenta que solo se presentaban 3 categorías (cédula de ciudadanía, cédula de extranjería y pasaporte), pero solo estaba presente 1 registro para la categoría de cédula de extranjería y pasaporte respectivamente. De igual manera se realizó el análisis del atributo ciudad y se tomó la decisión de prescindir de este, teniendo en cuenta que el estudiante durante su etapa académica por lo regular su lugar de residencia corresponde a la ciudad de Cali o sus alrededores. Los valores faltantes de la fecha y lugar de nacimiento fueron reemplazados por la moda.

En el caso de los valores de matrícula, se identificó la aplicación del método de sustitución por la media. En el caso de las becas/descuentos y los créditos financieros se pueden observar una cantidad considerable de datos faltantes (aproximadamente cifras cercanas a un 90% del total de los datos) debido a que representan datos que se estiman con valores a 0, teniendo en cuenta que estos casos son aquellos en los que el estudiante no tuvo beca o crédito como apoyo o financiación de su carrera.

En la Tabla 4 se muestra la cantidad de datos faltantes por cada atributo en cada clase.

Tabla 4 - Datos faltantes por clase

	No continuidad a posgrados	Continuidad a posgrados	Total
Fecha Grado	0	0	0
Promedio Académico	0	0	0
Ciudad	1896	475	2371
Género	0	0	0
Fecha Nacimiento	1	0	1
Lugar Nacimiento	10	0	0
Tipo Identificación	0	0	0
Programa Académico	0	0	0
Movilidad	0	0	0

Consejerías	0	0	0
Distinciones	0	0	0
Créditos Académicos	0	0	0
Monitorias	0	0	0
Valor Matrícula	4	0	4
Becas/Descuentos	2750	525	3275
Beca Permanente	0	0	0
Créditos Financieros	2516	470	2986
Doble Programa	0	0	0
Pruebas Académicas	0	0	0
Semestres Matriculados	0	0	0
Materias Perdidas	0	0	0

En la Figura 7 se muestra la distribución de las variables numéricas. Algunos atributos siguen una distribución normal, mientras que otros atributos como el valor de las becas/descuentos, los créditos financieros, las pruebas académicas y las materias perdidas siguen una distribución logarítmica normal.

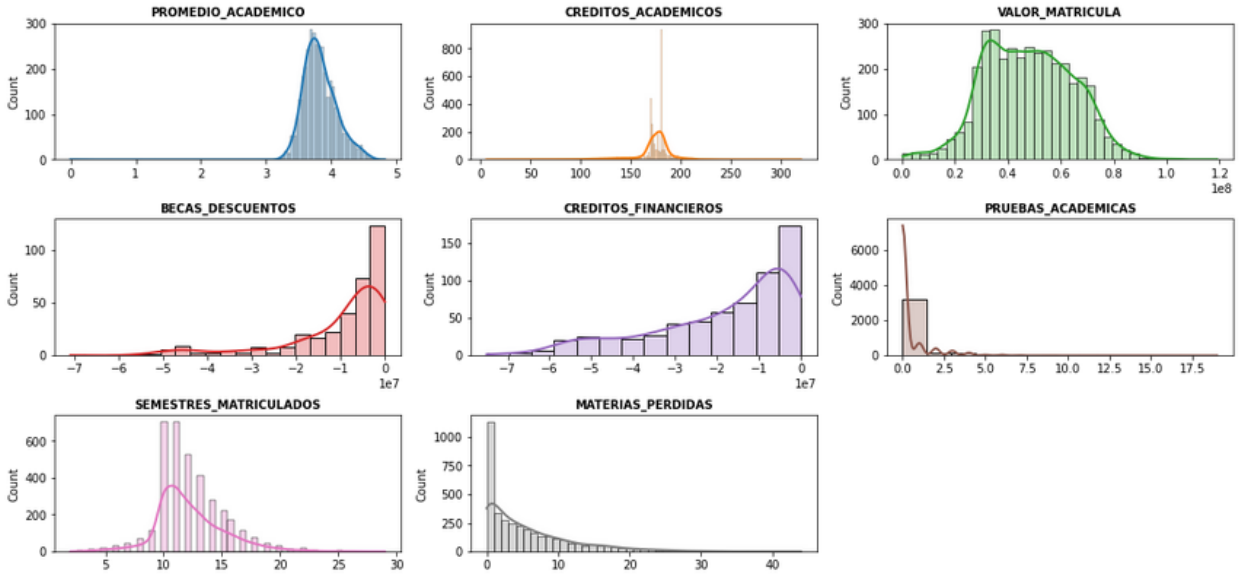


Fig. 7 Distribución de atributos numéricos

En la Figura 8 se muestra la matriz de correlación en donde se identifica que no hay un grado alto de relación lineal entre la mayoría de las variables. Al no estar altamente correlacionados estos atributos se logra inferir que estos miden diferentes características. Sin embargo, el análisis se centró en aquellos que tuvieron una correlación fuerte de manera directa o inversa: el promedio académico tiene una asociación en sentido inverso con las variables que determinan la cantidad de semestres matriculados y la cantidad de materias perdidas por lo que bajos promedios académicos pueden verse altamente influenciado con la duración del ciclo académico de un estudiante o de su rendimiento académico por las asignaturas perdidas. Por otro lado, pero siguiendo la misma línea del promedio académico, la asociación de las materias perdidas con la cantidad de pruebas académicas y la cantidad de semestres matriculados muestran que el bajo rendimiento académico de un estudiante que pierde materias puede alargar la duración en años de su ciclo académico en la universidad.

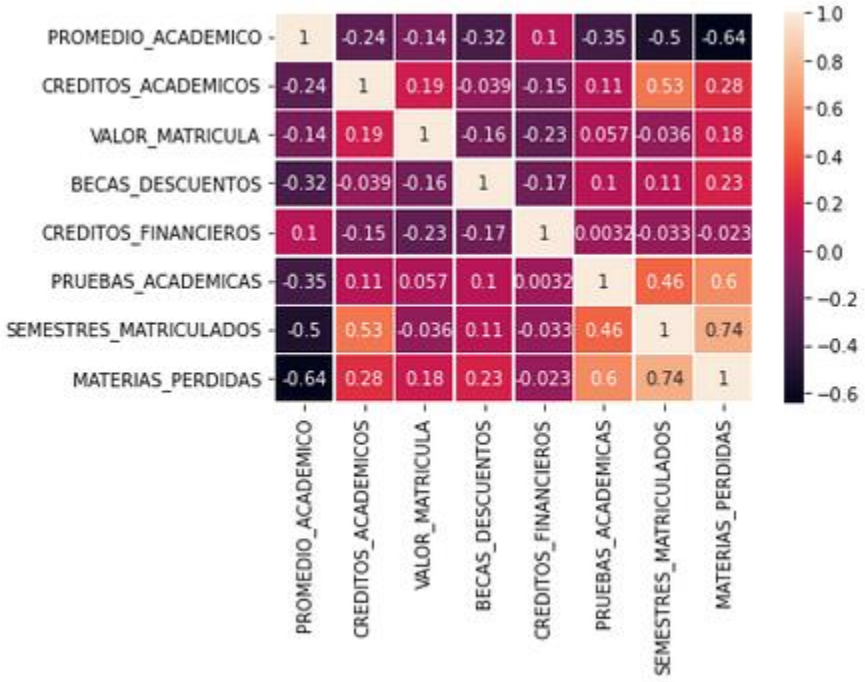


Fig. 8 Matriz de correlación

Para ampliar la información obtenida a partir de la matriz de correlación, en la Figura 9 y en la Figura 10 se muestra la dispersión de los datos, especialmente los créditos académicos de los egresados.

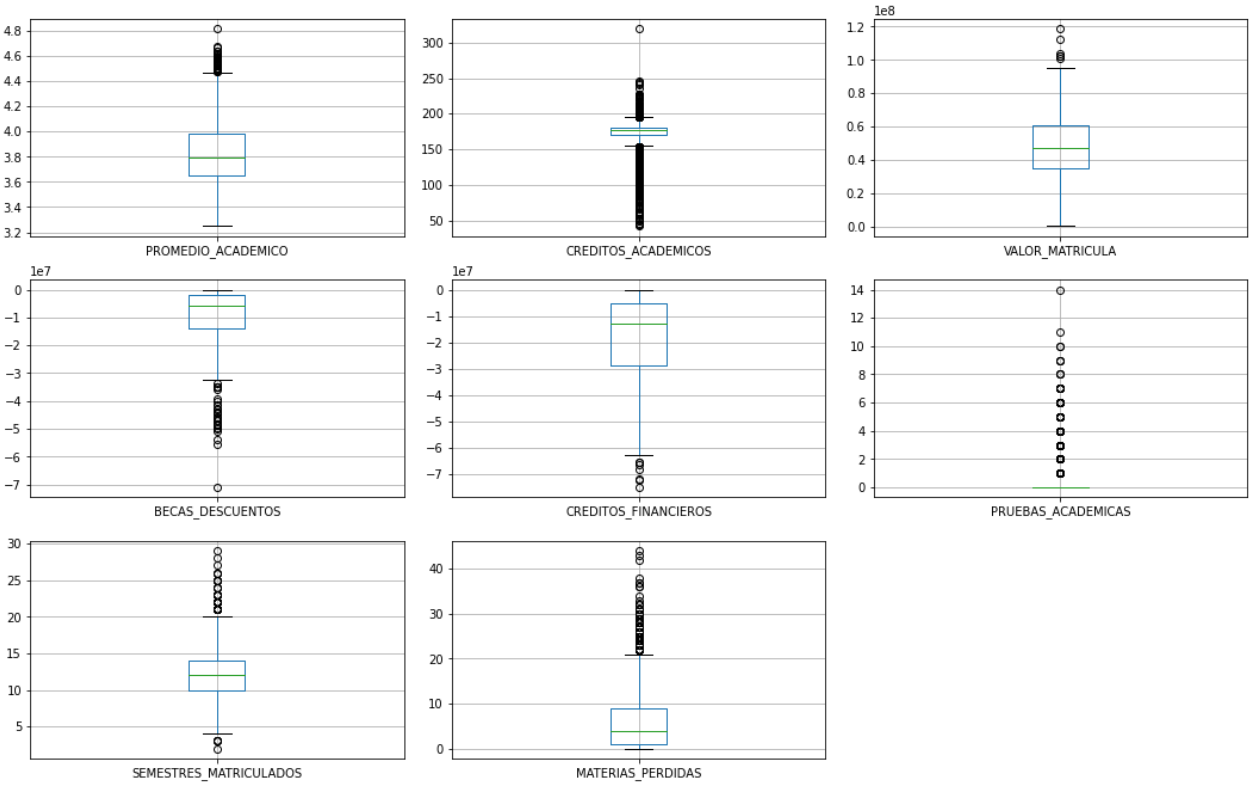


Fig. 9 Dispersión de datos

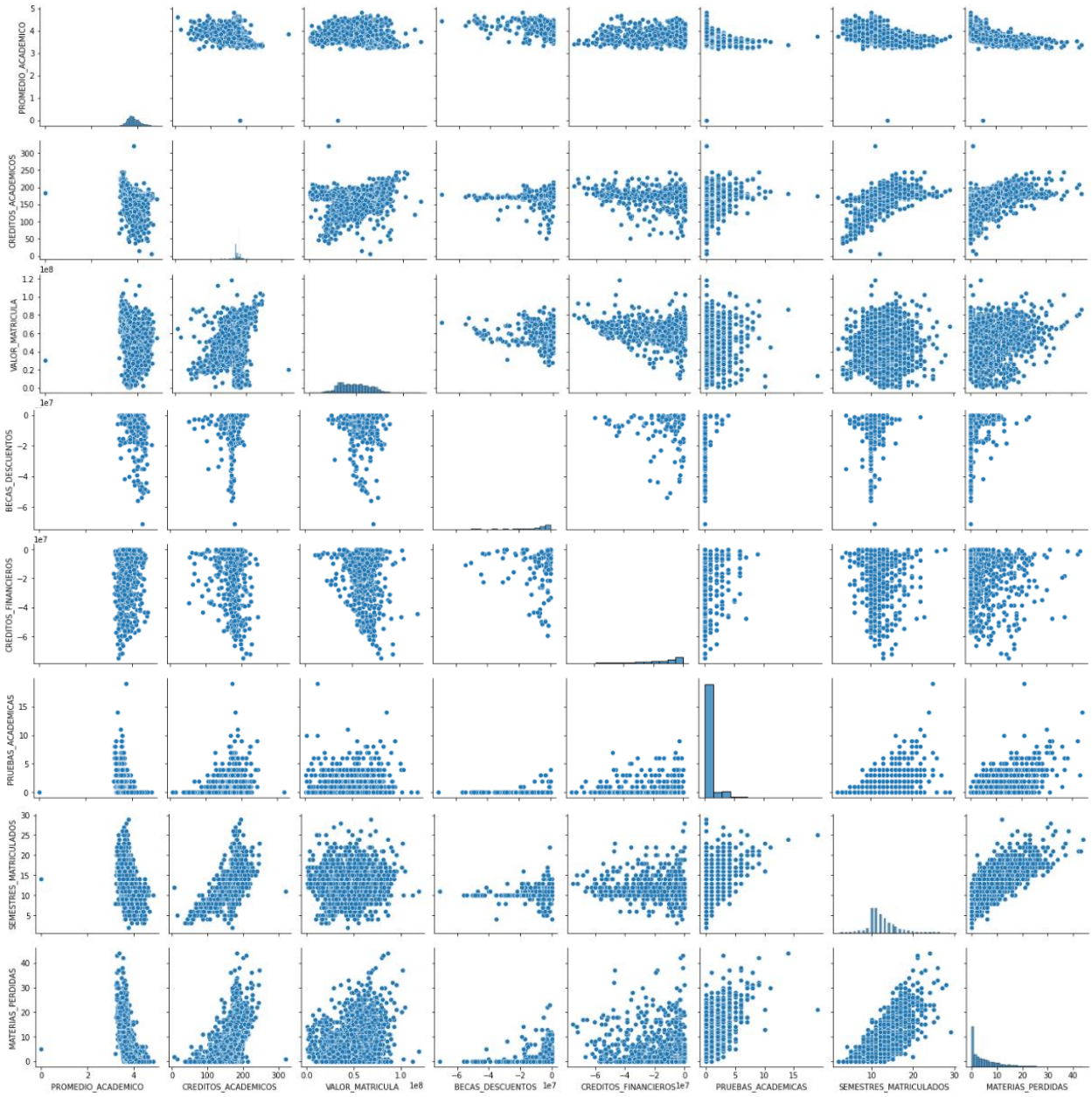


Fig. 10 Diagrama de pares

5.3. FASE III: PREPARACIÓN DE LOS DATOS

En esta fase se realizó la preparación de los datos con el objetivo de seleccionar los datos a los que se les va a aplicar las técnicas de modelado, se hizo una limpieza de estos datos para mejorar su calidad, se les dio el formato requerido y se generaron las variables adicionales requeridas para el procesamiento.

5.3.1 SELECCIÓN DE LOS DATOS. En esta etapa se realizó la selección de los datos y se determinó qué atributos se podían prescindir para la aplicación de las técnicas de modelado, teniendo en cuenta la importancia de dichos campos en relación con los objetivos de la minería de datos que se definieron en la fase de entendimiento del negocio. Atributos como el tipo de identificación y la ciudad son irrelevantes en el modelo, tal como se mencionó en la etapa de exploración de los datos. Para el caso de la ciudad de residencia se tomó la decisión de prescindir de esta característica teniendo en cuenta que 1130 estudiantes residían en Cali en el momento de realizar sus estudios de pregrado; se presentan algunos registros (74 estudiantes) en municipios cercanos pero la representación es baja como se puede observar en la Figura 11. Otros atributos irrelevantes para el modelo fueron el primer nombre, segundo nombre y apellido del estudiante que no fueron tenidos en cuenta y por lo tanto no se contemplaron en la etapa de la definición del conjunto de datos.

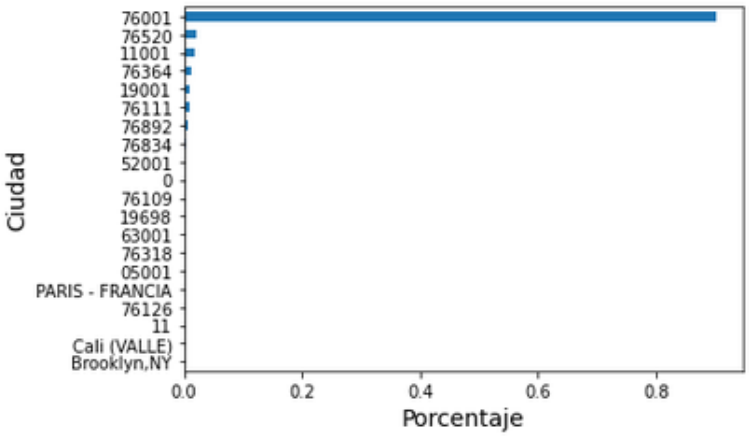


Fig. 11 Distribución de frecuencias de ciudades

5.3.2 LIMPIEZA DE LOS DATOS. En esta etapa se llevó a cabo la limpieza de los datos realizando el análisis de los datos faltantes o perdidos e identificando la ausencia de estos datos. Aunque la mayoría de los datos ya tenían una limpieza inicial debido a los procesos de BI de la universidad a través de los diferentes ETL (proceso de extracción, transformación y carga o centralización de datos) para el procesamiento de datos, se realizó una limpieza inicial tal como se menciona en la etapa de descripción de los datos eliminando registros mencionados en dichas etapas iniciales.

5.3.3 TRANSFORMACIÓN DE LOS DATOS. En esta etapa se realizaron las transformaciones de los campos categóricos a escala numérica (datos como el lugar de nacimiento, género y aquellos que son representados de forma cualitativa dicotómica) para el entrenamiento de los modelos y se realizó la transformación de los valores que eran cambiantes año tras año (por ejemplo los valores de matrícula suelen tener un incremento porcentual cada año, así como a presentar una variación

para la cohorte de admisión del estudiante en el programa académico lo que genera un impacto en el valor de las becas y créditos financieros) con los valores relacionados a la matrícula financiera.

Para el caso del lugar de nacimiento del estudiante, se reemplazaron los valores categóricos a numéricos según el tipo de codificación de tablas de municipios colombianos del DANE (entidad responsable de la planeación, levantamiento, procesamiento, análisis y difusión de las estadísticas oficiales de Colombia) y con valores de la moda para los datos que corresponden a ciudades extranjeras. En este caso se realizó la codificación de enteros porque si bien es cierto que no existe una relación ordinal entre las categorías, lo que puede dar como resultado un rendimiento deficiente o resultados inesperados en el modelo, el número de características resultantes aplicando One Hot Encoding puede desencadenar en una expansión masiva en el espacio de características (114 características adicionales), generando un ruido para el cual el modelo de aprendizaje puede tener un rendimiento extremadamente bajo, tardar más tiempo en entrenarse y tener asignación de recursos innecesarios para estas características adicionales.

Para el caso de las fechas de nacimiento y fechas de graduación se mantuvieron en esta etapa con sus valores iniciales para poder realizar integración de datos con el objetivo de obtener la edad del estudiante a la fecha de graduación.

Para el caso de los programas académicos se aplicó el método One Hot Encoding para codificar los valores categóricos como una matriz de vector binario, aprovechando la información contenida en un valor de categoría sin la confusión causada por la ordinalidad que puede representar al aplicar codificación de enteros. La estrategia que se implementó con este método fue crear una columna para cada programa y, para cada registro, marcar con un 1 el programa al que pertenece el estudiante y dejar las demás con 0. La representación se puede ver reflejada en la Tabla 5.

Tabla 5 - Método One Hot Encoding para programas

PROGRAMA_ACADEMICO	BIOLOGIA	ING_CIVIL	ING_ELECTRONICA	ING_INDUSTRIAL	ING_SISTEMAS	MAT_APLICADAS
Ingeniería Civil	0.00	1.00	0.00	0.00	0.00	0.00
Ingeniería Industrial	0.00	0.00	0.00	1.00	0.00	0.00
Ingeniería Industrial	0.00	0.00	0.00	1.00	0.00	0.00
Ingeniería de Sistemas	0.00	0.00	0.00	0.00	1.00	0.00
Ingeniería Industrial	0.00	0.00	0.00	1.00	0.00	0.00
...	...	...	...	...	...	...
Ingeniería Civil	0.00	1.00	0.00	0.00	0.00	0.00
Ingeniería Electrónica	0.00	0.00	1.00	0.00	0.00	0.00
Ingeniería de Sistemas	0.00	0.00	0.00	0.00	1.00	0.00
Ingeniería Industrial	0.00	0.00	0.00	1.00	0.00	0.00
Ingeniería Electrónica	0.00	0.00	1.00	0.00	0.00	0.00

Para el caso de los valores asociados a la matrícula financiera del estudiante, se optó por utilizar el valor de la matrícula como base para calcular el valor porcentual de los otros valores respecto a dicho valor, teniendo en cuenta que el valor de la matrícula es un valor cambiante a través del tiempo. Para este caso se dividió el valor de las becas/descuentos y los créditos financieros por el valor de la matrícula para obtener el porcentaje. Al final se elimina el atributo de valor de matrícula.

Finalmente, se convirtieron en valores numéricos los registros de las clases de salida que determinan la continuidad de un egresado o no en estudios de posgrados en la universidad.

5.3.4 INTEGRACIÓN DE LOS DATOS. En esta etapa se realizó la integración de datos para la creación de nuevos campos en el set de datos. Se integraron los atributos de la fecha de nacimiento del estudiante y la fecha de egreso para obtener la edad del estudiante al momento de graduarse de su programa académico en la universidad.

En la Figura 12 se observa la distribución de los datos referentes a la edad representados en un histograma en donde se puede observar que las edades de los estudiantes tienden a seguir una distribución logarítmica normal.

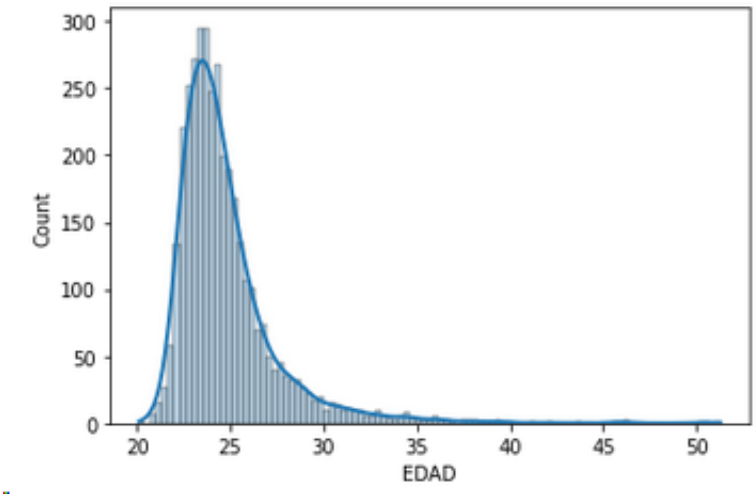


Fig. 12 Histograma de distribución de edades

5.3.5 FORMATEO DE LOS DATOS. En esta etapa se realizó la discretización o binning de los atributos referentes al promedio académico y la cantidad de créditos académicos. El objetivo del proceso fue reducir la cantidad de valores para los atributos continuos, dividiendo el rango en intervalos, por lo tanto, se aplicó especialmente para los atributos anteriormente mencionados que podían ser descritos en forma de intervalos. Posteriormente se seleccionaron los números de intervalos para cada atributo basado en las medidas de centralidad como el valor mínimo y el valor máximo del atributo.

Para el caso del atributo del promedio académico, se discretizó en 5 intervalos de diferentes promedios: muy bajo, bajo, medio, alto y muy alto. En el caso del número de créditos académicos la discretización se realizó en 3 intervalos de diferentes número de créditos: poco, normal y mucho.

En la Tabla 6 se pueden observar los atributos que se discretizaron, así como la información asociada al número de intervalos que se definió para aplicar las herramientas de discretización.

Tabla 6 - Discretización de atributos

	Promedio Académico	Créditos Académicos
min	3.25	42.00
max	4.82	320.00
counts	589	158
bins	5	3

5.4. FASE IV: MODELADO

En esta fase se seleccionaron 3 técnicas de modelado basada en la literatura de referencia[18] que definen a los algoritmos basados en árboles como los mejores y más utilizados métodos de aprendizaje automático y cuya selección se vio motivada principalmente con la necesidad de hacer un análisis interpretativo y encontrar características más relacionadas con la decisión de continuar o no el posgrado. Dentro de las principales ventajas de trabajar con modelos basados en árboles están que no son paramétricos (son excelentes cuando tenemos muchos datos), interpretables (lo que significa que después de construir el modelo, podemos interpretar el modelo, no solo las predicciones), permiten reducir la dimensionalidad, en la mayoría de casos son rápidos al ser simples y el preprocesamiento de los datos es más fácil porque no se hace necesario escalar los datos. A partir de lo mencionado anteriormente, las técnicas seleccionadas fueron los árboles de decisión (Decision Trees), los bosques aleatorios (Random Forests) y potenciación del gradiente (Gradient Boosting) que son técnicas basadas en árboles, de uso frecuente por su sencilla implementación y fácil interpretación. Los métodos basados en árboles engloban a un conjunto de técnicas supervisadas no paramétricas (no hay suposición acerca del espacio de distribución y la estructura del clasificador) que consiguen segmentar el espacio de los predictores en regiones simples, dentro de las cuales es más sencillo manejar las interacciones[19]. Es esta última característica la que les proporciona gran parte de su potencial.

Los árboles de decisión son estructuras de posibles soluciones a una decisión basadas en ciertas condiciones, es uno de los algoritmos de aprendizaje supervisado más utilizados en machine learning y pueden realizar tareas de clasificación o regresión.

Los bosques aleatorios están formados por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento. Esto implica que cada árbol se entrena con unos datos ligeramente distintos. En cada árbol individual, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

Un modelo Gradient Boosting está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

5.4.1 PROCESO DE ENTRENAMIENTO Y OPTIMIZACIÓN. Los algoritmos de aprendizaje automático aprenden de los datos con los que los entrenamos. A partir de ellos, intentan encontrar o inferir el patrón que les permita predecir el resultado para un nuevo caso. Pero, para poder validar si un modelo funciona, necesitaremos probarlo con un conjunto de datos diferente[19]. Por ello, en todo proceso de aprendizaje automático, los datos de trabajo se dividen

en dos partes: datos de entrenamiento y datos de prueba o test.

Antes de dar inicio a la construcción de los diferentes modelos, se realizó la separación del set de datos en datos de entrenamiento (son los datos usados para entrenar un modelo) y datos de prueba (son los datos reservados para comprobar si el modelo generado a partir de los datos de entrenamiento funciona). Adicionalmente, se tuvo en cuenta la importancia que el conjunto de datos de prueba tuviera un volumen suficiente como para generar resultados significativos, y a la vez, que fueran representativos del conjunto de datos global, por lo que se decidió repartir el conjunto de datos en un 70% de datos de entrenamiento y un 30% de datos de test con el objetivo de usar la razón de error como medida de calidad. El set de entrenamiento logró tener clases en una proporción de 2108 y 374 de registros de estudiantes que no realizarían posgrados y aquellos que sí lo harían, respectivamente.

Posteriormente, se procedió a generar las diferentes instancias de los modelos con los valores por defecto de cada una de las técnicas según scikit-learn (la biblioteca para aprendizaje automático usado en el proyecto).

En el caso de los árboles de decisión algunos de los parámetros más importantes son aquellos que detienen el crecimiento del árbol, entre los cuales se encuentran[20]:

- **max_depth**: profundidad máxima que puede alcanzar el árbol.
- **min_samples_split**: número mínimo de observaciones que debe de tener un nodo para que pueda dividirse.
- **min_samples_leaf**: número mínimo de observaciones que debe de tener cada uno de los nodos hijos para que se produzca la división.
- **max_leaf_nodes**: número máximo de nodos terminales.
- **random_state**: semilla para que los resultados sean reproducibles.

Para el caso de los bosques aleatorios algunos de los parámetros más importantes son aquellos que detienen el crecimiento de los árboles, los que controlan el número de árboles y predictores incluidos, y los que gestionan la paralelización, entre los cuales se encuentran[20]:

- **n_estimators**: número de árboles incluidos en el modelo.
- **max_depth**: profundidad máxima que pueden alcanzar los árboles.
- **min_samples_split**: número mínimo de observaciones que debe de tener un nodo para que pueda dividirse.
- **min_samples_leaf**: número mínimo de observaciones que debe de tener cada uno de los nodos hijos para que se produzca la división.
- **max_leaf_nodes**: número máximo de nodos terminales que pueden tener los árboles.
- **max_features**: número de predictores considerados a en cada división.

- **n_jobs**: número de cores empleados para el entrenamiento. En random forest los árboles se ajustan de forma independiente, por lo que la paralelización reduce notablemente el tiempo de entrenamiento.
- **random_state**: semilla para que los resultados sean reproducibles.

Para el caso de la potenciación del gradiente algunos de los parámetros más importantes son aquellos que controlan el crecimiento de los árboles, la velocidad de aprendizaje del modelo, y los que gestionan la parada temprana para evitar overfitting, entre los cuales se encuentran[20]:

- **learning_rate**: reduce la contribución de cada árbol multiplicando su influencia original por este valor.
- **n_estimators**: número de árboles incluidos en el modelo.
- **max_depth**: profundidad máxima que pueden alcanzar los árboles.
- **min_samples_split**: número mínimo de observaciones que debe de tener un nodo para que pueda dividirse.
- **min_samples_leaf**: número mínimo de observaciones que debe de tener cada uno de los nodos hijos para que se produzca la división.
- **max_leaf_nodes**: número máximo de nodos terminales que pueden tener los árboles.
- **max_features**: número de predictores considerados a en cada división.
- **subsample**: proporción de observaciones utilizadas para el ajuste de cada árbol.
- **validation_fraction**: proporción de datos separados del conjunto entrenamiento y empleados como conjunto de validación para determinar la parada temprana (early stopping).
- **n_iter_no_change**: número de iteraciones consecutivas en las que no se debe superar la tolerancia para que el algoritmo se detenga (early stopping).
- **tol**: porcentaje mínimo de mejora entre dos iteraciones consecutivas por debajo del cual se considera que el modelo no ha mejorado.
- **random_state**: semilla para que los resultados sean reproducibles.

Paso a seguir se entrenaron los modelos evaluando la capacidad predictiva inicial y utilizando los parámetros por defecto de la biblioteca scikit-learn como valores base, obteniendo las diferentes métricas de evaluación del rendimiento que se pueden observar en la Tabla 7. Estas métricas de evaluación permiten valorar el rendimiento de un modelo de aprendizaje automático, que es un componente integral de cualquier proyecto de ciencia de datos. Su objetivo es estimar la precisión de la generalización de un modelo sobre los datos futuros (no vistos/fuera de muestra). Las métricas utilizadas fueron *accuracy* (nos indica de todas las clases cuántas se predijeron correctamente), *precision* (nos indica de todas las clases positivas que se han predicho cuántas son realmente positivas), *recall* (nos indica de todas las clases positivas cuántas se predijo correctamente) y *f1* (permite comparar dos modelos de baja precisión y alta exhaustividad, usa la media armónica para castigar los valores extremos).

Tabla 7 - Métricas de evaluación con modelos base

Métrica	Decision Tree	Random Forests	Gradient Boosting
Accuracy	98.6854%	99.2488%	99.1549%
Precision	96.7948%	98.7179%	98.7096%
Recall	94.375%	96.25%	95.625%
F1	95.5696%	97.4683%	97.1428%

5.4.1.1. Tunning de árboles de decisión. A partir de la generación de los modelos con los parámetros por defecto, se procedió a realizar el tuning de los parámetros. Encontrar el valor óptimo de profundidad máxima que puede alcanzar el árbol es una forma de ajustar el modelo con la técnica de árboles de decisión. En las Figuras 13 y 14 se muestra que la precisión media de los árboles de decisión con diferentes valores para el parámetro de *max_depth* es cuando el parámetro se evalúa en 3 con un score de 99,530%. Es importante tener en cuenta que *max_depth* no es lo mismo que la profundidad de un árbol de decisión; *max_depth* es una forma de podar previamente un árbol de decisión.

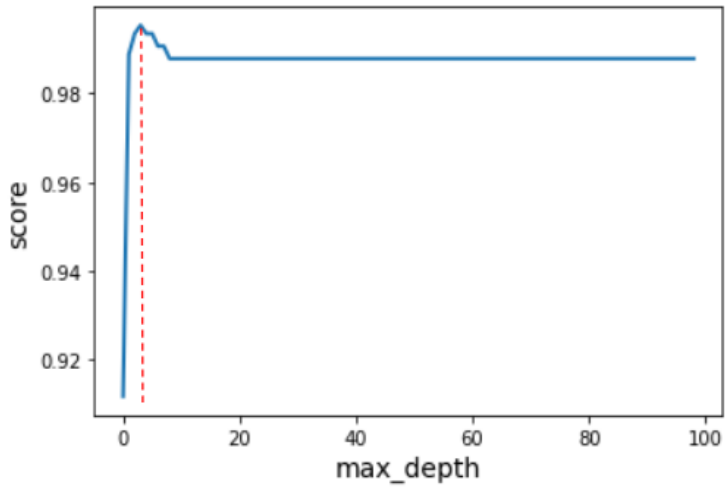


Fig. 13 Precisión media para diferentes valores de max_depth (rango 1 a 100) en árboles de decisión

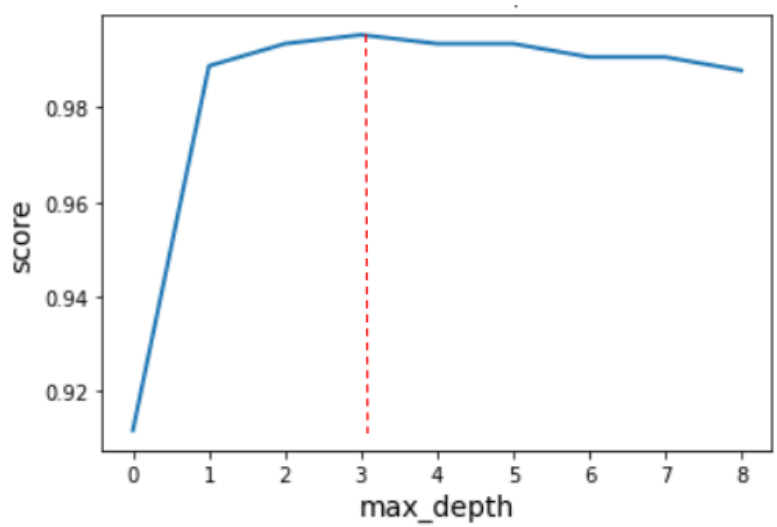


Fig. 14 Precisión media para diferentes valores de max_depth (rango 1 a 10) en árboles de decisión

Otro de los aspectos a tener en cuenta es el peso de las clases, debido a lo anterior se definió el parámetro *classs_weight="balanced"*, con esto el algoritmo se encargó de equilibrar a la clase minoritaria durante el entrenamiento; básicamente significa replicar la clase más pequeña hasta tener tantas muestras como en la clase mayoritaria, pero de forma implícita. Teniendo en cuenta el ajuste del modelo inicial haciendo tunning de los parámetros *max_depth* y *class_weigth* se obtuvieron las siguientes métricas al evaluar el modelo con los datos de validación: *accuracy* de 98.2159%, *precision* de 90.7514%, *recall* de 98.125% y *f1* de 94.294%.

Se experimentó hacer tunning de otros hiperparámetros, pero los resultados obtenidos presentaron métricas que no resultaron tan óptimas. El experimento consistió en incrementar los valores (4, 10, 20, 40) de *min_samples_split* (por defecto el parámetro es igual a 2) y de *min_samples_leaf* (por defecto el parámetro es igual a 1) basados en un estudio empírico sobre el ajuste de hiperparámetros de los árboles de decisión[22], en donde se indica que los valores ideales de *min_samples_split* tienden a estar entre 1 y 40 para el algoritmo CART, que es el algoritmo implementado en scikit-learn, mientras que los valores ideales de *min_samples_leaf* tienden a estar entre 1 y 20 para el algoritmo CART, que es el algoritmo implementado en scikit-learn, obteniendo las siguientes métricas combinadas con el tuning de las métricas *max_depth* y *class_weight*: *accuracy* de 98.1220%, *precision* de 90.6976%, *recall* de 97.5% y *f1* de 93.9759%.

El proceso de construcción del modelo se ajustaba muy bien a las observaciones empleadas como entrenamiento. A partir de ello, se realizó el proceso de validación entre los errores de entrenamiento y test a fin de determinar un posible overfitting que reduce la capacidad predictiva del modelo al aplicarlo a nuevos datos. La razón de este comportamiento radica en la facilidad con la que los árboles se ramifican adquiriendo estructuras complejas. De hecho, si no se limitan las divisiones, todo árbol termina ajustándose perfectamente a las observaciones de

entrenamiento creando un nodo terminal por observación. El error de entrenamiento (98.4286%) arrojó un valor mayor al del test (98.2159%) lo cual no se interpretaría como overfitting porque la diferencia es muy pequeña.

Como experimento se tomó la decisión de aplicar la estrategia de pruning por encima de la parada temprana (early stopping). La estrategia de controlar el tamaño del árbol mediante reglas de parada tiene un inconveniente, el árbol se crece seleccionando la mejor división en cada momento. Al evaluar las divisiones sin tener en cuenta las que vendrán después, nunca se elige la opción que resulta en el mejor árbol final, a no ser que también sea la que genera en ese momento la mejor división. A este tipo de estrategias se les conoce como greedy.

Otro experimento fue generar una alternativa no greedy, que consiste en generar árboles grandes, sin condiciones de parada más allá de las necesarias por las limitaciones computacionales, y después podarlos (pruning), manteniendo únicamente la estructura robusta que consigue un *evaluation_error* bajo. La selección del sub-árbol óptimo puede hacerse mediante cross-validation, sin embargo, dado que los árboles se crecen lo máximo posible (tienen muchos nodos terminales) no suele ser viable estimar el test error de todas las posibles sub-estructuras que se pueden generar. En su lugar, se recurrió al cost complexity pruning por validación cruzada. En la Figura 15 (aunque no se puede observar con detalle) se encuentran los valores mínimos y máximos del parámetro *ccp_alpha* que corresponden a los valores entre 0.0 y 0.0242, obteniendo la máxima precisión de validación. Aunque la precisión de entrenamiento disminuye a 0.99, en ese punto el modelo se volvió más generalizado y funciona mejor con datos no vistos.

Una vez identificado el valor óptimo de *ccp_alpha*, se reentrenó el árbol indicando este valor en sus argumentos obteniendo una precisión de validación igual a 98.9671%. Con estos ajustes se obtuvieron las siguientes métricas al evaluar el modelo: *accuracy* de 98.9671%, *precision* de 96.2732%, *recall* de 96.8750% y *f1* de 96.5732%.

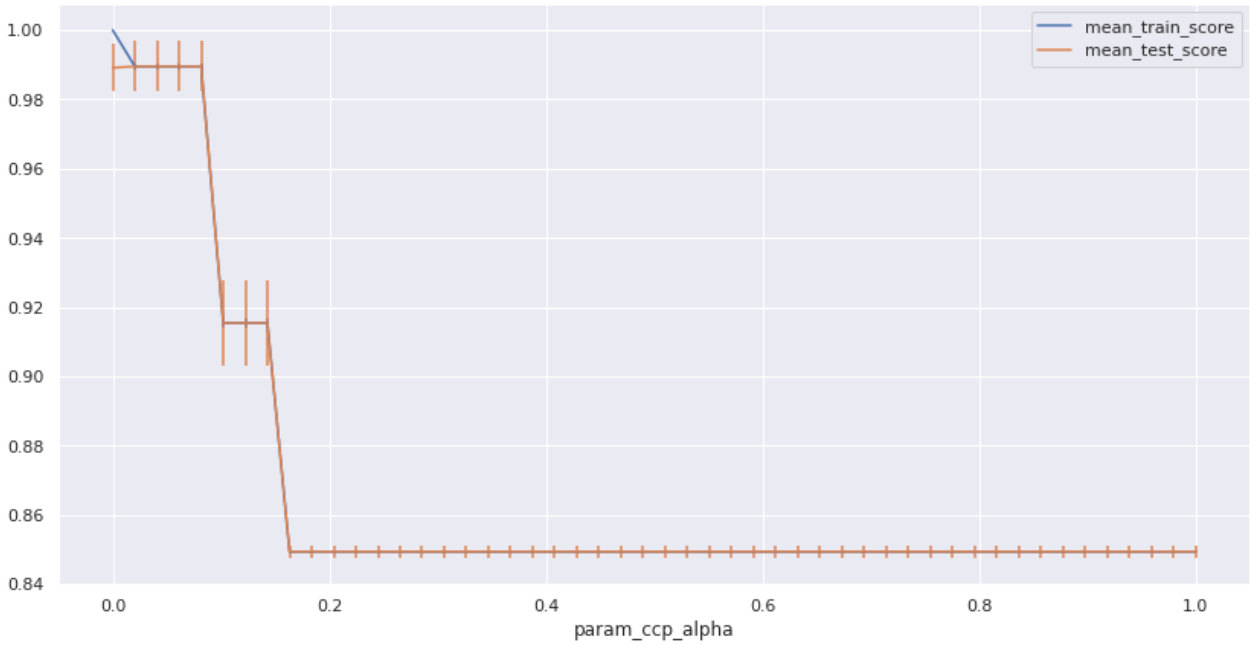


Fig. 15 Error de validación cruzada vs hiperparámetro ccp_alpha

5.4.1.2. Tunning de bosques aleatorios. En el caso de la técnica de bosques aleatorios, al igual que con la técnica de árboles de decisión, se procedió a encontrar el valor óptimo de profundidad máxima que puede alcanzar un árbol aplicando tuning al modelo. Se procedió a realizar un análisis con un rango de 1 a 100 para encontrar el valor óptimo y una vez identificado dicho valor, se procedió a realizar el mismo ejercicio, pero esta vez en un rango de 1 a 20 para conocer con mayor precisión el valor óptimo. En las Figuras 16 y 17 se puede apreciar que la precisión media con diferentes valores para el parámetro de *max_depth* es cuando el parámetro se evalúa en 12 con un score de 99,3427%.

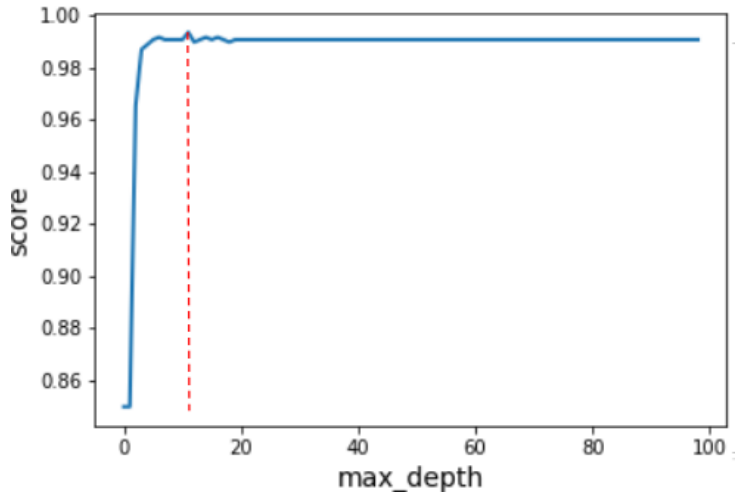


Fig. 16 Precisión media para diferentes valores de max_depth (rango 1 a 100) en bosques aleatorios

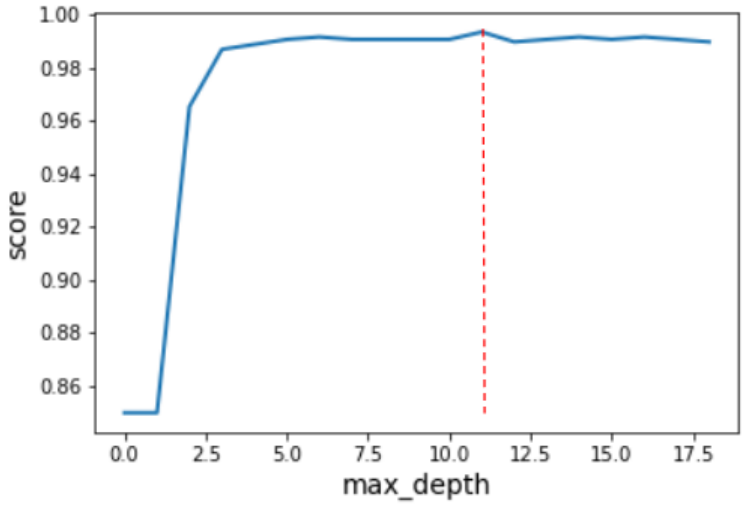


Fig. 17 Precisión media para diferentes valores de max_depth (rango 1 a 20) en bosques aleatorios

Al igual que con la técnica de árboles de decisión, para los bosques aleatorios también se definió el parámetro *classs_weight="balanced"*. Adicionalmente, a través de la curva de validación (validation curve), la cual es una buena forma de verificar visualmente los valores potencialmente optimizados de los hiperparámetros del modelo, se logró encontrar el valor óptimo para el parámetro de *n_estimators*, el cuál indica el número de árboles de decisión que se construyen en el modelo de random forest, obteniendo un valor de *n_estimatos*=150 como se puede visualizar en la Figura 18; esta curva de validación se creó con los valores [10, 20, 50, 100, 150, 300] como los diferentes valores a probar para *n_estimators*. Con estos ajustes se obtuvieron las siguientes métricas al evaluar el modelo: *accuracy* de 99.0610%, *precision* de 97.4683%, *recall* de 96.25% y *f1* de 96.8553%.

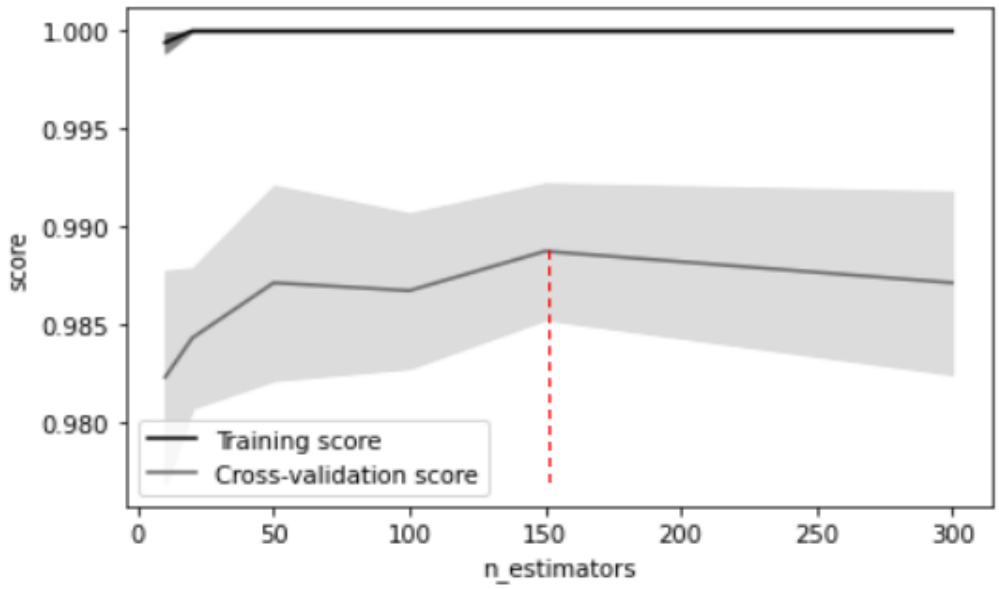


Fig. 18 Precisión media para diferentes valores de n_estimators en bosques aleatorios

Si bien RandomForestClassifier tiene valores por defecto para sus hiperparámetros, no se puede saber de antemano si estos son los más adecuados, la forma de identificarlos es mediante el uso de estrategias de validación, por ejemplo validación cruzada. Los modelos de bosques aleatorios tienen la ventaja de disponer del Out-of-Bag error, lo que permite obtener una estimación del error medio de predicción en cada muestra de entrenamiento usando solo los árboles no tenidos en cuenta en su muestra de arranque, sin recurrir a la validación cruzada que es computacionalmente costosa[20], sin embargo, se aplicó validación cruzada para realizar la comparación de los hiperparámetros óptimos para el modelo. Se aplicó out-of-bag error para hacer tuning de hiperparámetros con datos de los 2 *n_estimators* (50, 150) más altos que se obtuvieron en la curva de validación, las 22 características del modelo, los 4 valores más altos de los valores de *max_depth* (7, 12, 15, 17) y, los criterios *gini* y *entropy*. Los resultados de los valores óptimos se pueden observar en la Tabla 8. De igual manera, en la Tabla 9 se pueden observar los resultados obtenidos en la aplicación de la validación cruzada con el mismo rango de valores utilizado para el out-of-bag error.

Tabla 8 - Grid Search basado en out-of-bag score en random forests

oob_accuracy	criterion	max_depth	max_features	n_estimators
99,2345%	entropy	17.0	17	50
99,2345%	entropy	NaN	17	50
99,2345%	entropy	15.0	17	50
99,2345%	entropy	12.0	17	50
99,1942%	entropy	7.0	21	50
99,1942%	entropy	7.0	18	50
99,1942%	entropy	7.0	17	50
99,1539%	entropy	12.0	18	50
99,1539%	gini	17.0	15	50
99,1539%	entropy	7.0	20	150

Tabla 9 - Grid Search basado en validación cruzada en random forests

criterion	max_depth	max_features	n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score
entropy	7	11	50	99,1138%	0,3350%	99,8590%	0,0642%
gini	7	10	50	99,1137%	0,2730%	99,8691%	0,0604%
entropy	7	10	50	99,1004%	0,3273%	99,8623%	0,0566%
entropy	17	11	50	99,1003%	0,3436%	99,9966%	0,0126%

Los resultados obtenidos con ambas técnicas de tuning se pueden observar en la Tabla 10.

Tabla 10 - Grid Search basado en out-of-bag score vs cross-validation en random forests

métrica	out-of-bag error	cross-validation
Accuracy	99,2488%	99,2488%
Precision	98,7179%	99,3506%
Recall	96,2500%	95,6250%
F1	97,4683%	97,4522%

5.4.1.3. Tunning de potenciación del gradiente. En Gradient Boosting, el número de árboles es un hiperparámetro crítico en cuanto que, conforme se añaden árboles, se incrementa el riesgo de overfitting, por lo tanto, no se ejecutaron estrategias para obtener el valor óptimo de *n_estimators* por lo que su optimización se realizó con validación cruzada. Junto con el número de árboles, el *learning_rate* es el hiperparámetro más importantes en Gradient Boosting, ya que es el que permite controlar cómo de rápido aprende el modelo y con ello el riesgo de llegar al overfitting. Estos dos hiperparámetros son interdependientes, cuanto menor es el *learning_rate*, más árboles se necesitan para alcanzar buenos resultados pero menor es el riesgo de overfitting[20]. Los valores estimados por validación cruzada indican que, el mejor modelo se consigue con un *learning_rate* de 0.1. La profundidad de los árboles (*max_depth*) en los modelos Gradient Boosting suele ser un valor muy bajo, haciendo así que cada árbol solo pueda aprender una pequeña parte de la relación entre predictores y variable respuesta (weak learner).

En la búsqueda óptima de hiperparámetros en potenciación del gradiente, se empleó la estrategia de no incluir el número de árboles como hiperparámetro dentro de los parámetros a optimizar por validación cruzada. En su lugar, se utilizó por defecto un número elevado activando la parada temprana para evitar el overfitting. Se aplicó validación cruzada con datos de *n_estimators* (50, 100, 500, 1000) y valores de *learning_rate* de 0.001, 0.01 y 0.1. Los resultados de los valores óptimos se pueden observar en la Tabla 11.

Tabla 11 - Grid Search basado en validación cruzada en potenciación del gradiente

learning_rate	max_depth	max_features	n_estimators	subsample	mean_test_score	std_test_score	mean_train_score	std_train_score
0,01	3	auto	1000	0,5	99,3553%	0,0572%	99,9597%	0,0285%
0,1	3	auto	100	1	99,2749%	0,1970%	99,9396%	0,0493%
0,1	5	auto	100	1	99,2346%	0,1503%	100,0000%	0,0000%
0,01	5	auto	1000	1	99,2345%	0,0565%	100,0000%	0,0000%

5.4.2 RESULTADOS Y ANÁLISIS. Si bien la preparación de los datos y el entrenamiento de un modelo de aprendizaje de máquina es un paso clave en el proceso de aprendizaje automático, es igualmente importante medir el rendimiento del modelo entrenado. Al utilizar diferentes métricas para la evaluación del rendimiento, debemos estar en posición de mejorar el poder de predicción general de nuestro modelo sobre datos no vistos antes. Si no se realiza una evaluación adecuada del modelo aprendizaje automático utilizando diferentes métricas, y se usa sólo la precisión, puede darse un problema cuando el modelo respectivo se despliega sobre datos no vistos y puede dar lugar a malas predicciones. Esto sucede porque, en casos como éste, nuestros modelos no aprenden, sino que memorizan; por lo tanto, no pueden generalizar bien sobre datos no vistos.

A partir del tuning de los hiperparámetros de los diferentes modelos y su posterior evaluación, se obtuvieron las diferentes métricas (*accuracy, precision, recall y f1*) y se consolidaron los resultados de los diferentes modelos generados, haciendo un análisis de los valores obtenidos a partir de la matriz de confusión de los modelos, a fin de visualizar el rendimiento de los mismos y tener una mejor idea del rendimiento de cada modelo de aprendizaje automático. Adicionalmente, de cada modelo se logró identificar la importancia de los predictores en el modelo.

5.4.1.4. Árboles de decisión. En la Figura 19 se puede apreciar que con la configuración en donde se aplicó el primer ajuste al modelo generó un nodo raíz que hace una primera subdivisión por el atributo de becas y descuentos y las salidas van a la izquierda por aquellos estudiantes que no tuvieron becas y a derecha van aquellos estudiantes que sí tuvieron becas, así como sucede con la Figura 20 con la selección de parámetros finales en donde se obtuvo una profundidad máxima del árbol de 9. Es importante resaltar que en ambos casos el hecho de no tener beca es un indicador de que el estudiante no continuaría sus estudios de posgrado y, después de las becas, el siguiente atributo corresponde a los créditos financieros.

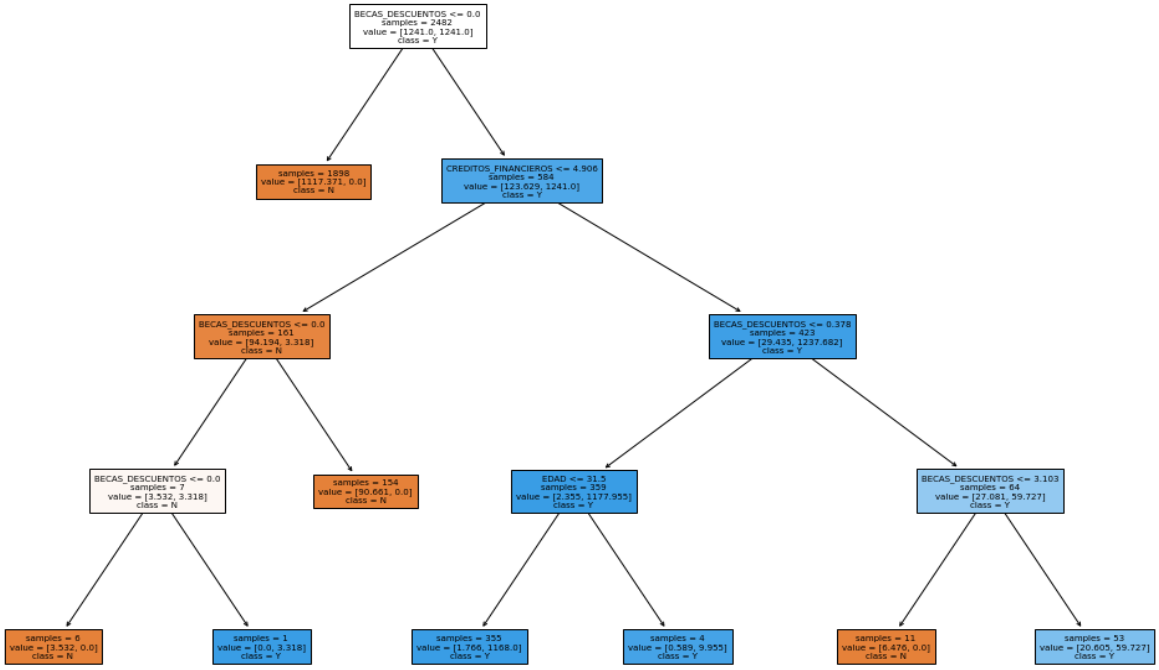


Fig. 19 Estructura del árbol de decisión de profundidad 4

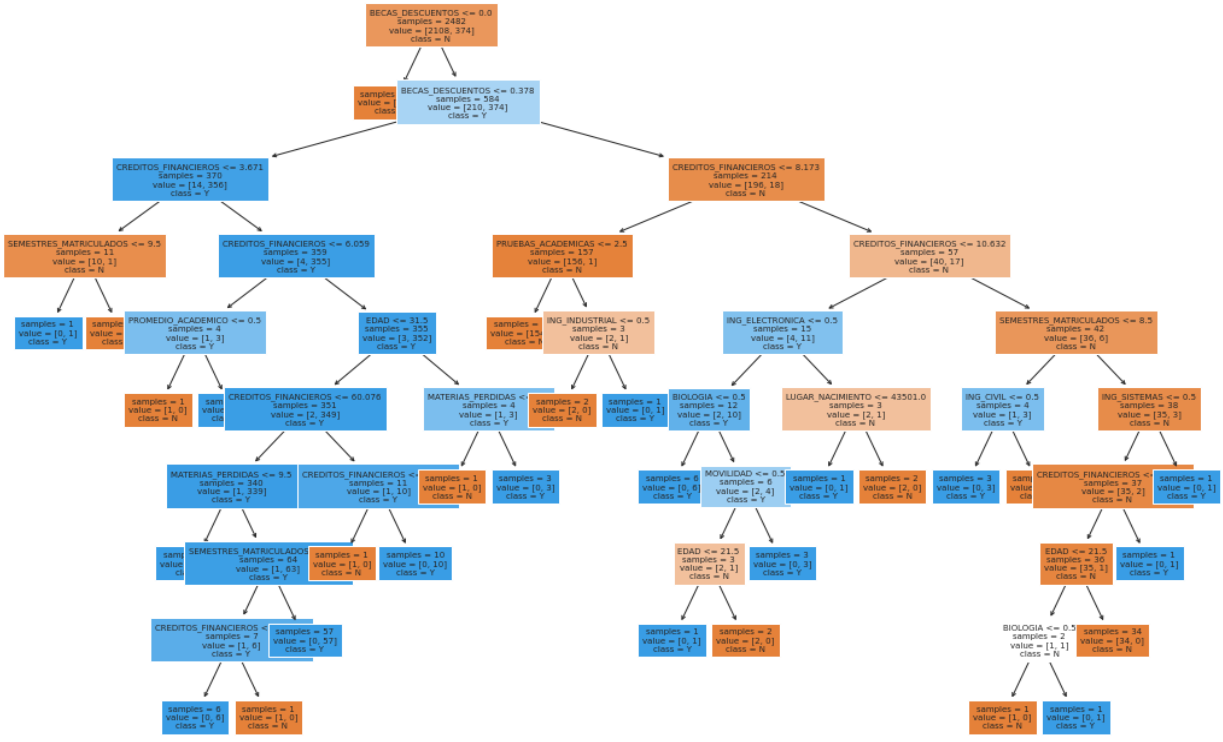


Fig. 20 Estructura del árbol de decisión con parámetros finales

La matriz de confusión es una matriz que permite visualizar el rendimiento de los modelos de aprendizaje automático de clasificación. Con esta visualización, se puede tener una mejor idea del rendimiento del modelo de aprendizaje automático. La matriz resultante permite representar los verdaderos positivos, los falsos positivos, los falsos negativos y los verdaderos negativos como se observa en la Figura 21 y 22. Si se realiza la comparación de las 2 figuras se puede observar que se presentó una disminución de los falsos positivos (una disminución de 18 falsos positivos que corresponde a un 1.5%)

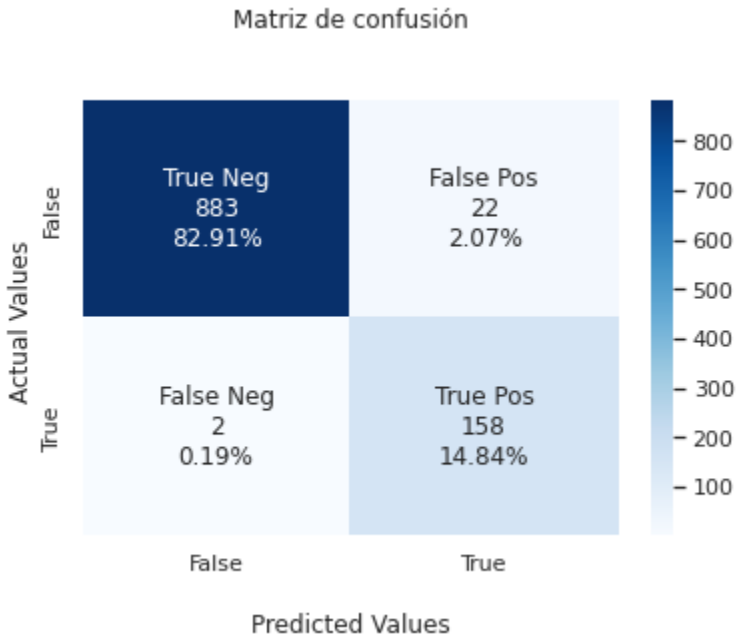


Fig. 21 Matriz de confusión de árboles de decisión sin tunning

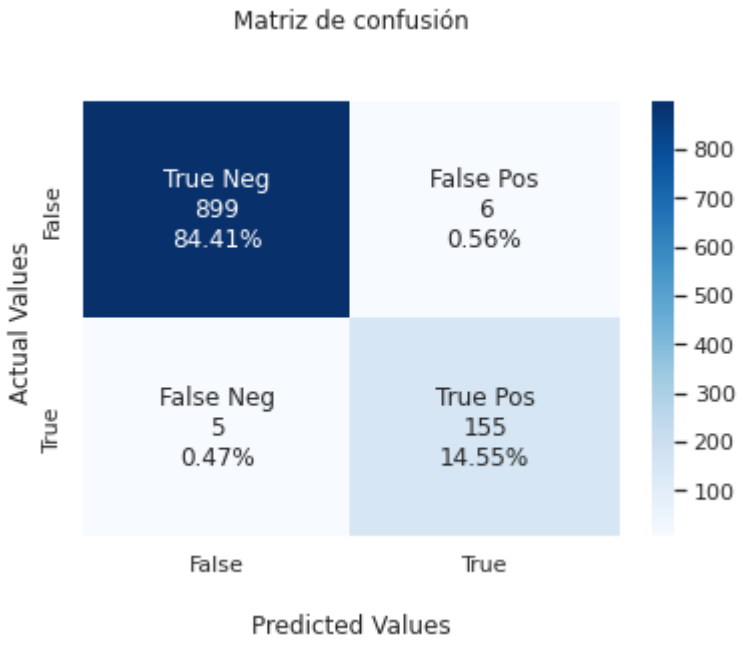


Fig. 22 Matriz de confusión de árboles de decisión final

La sensibilidad (o recall) representa la tasa de verdaderos positivos (True Positive Rate) ó TP. Es la proporción entre los casos positivos bien clasificados por el modelo, respecto al total de positivos. Permite representar la habilidad del modelo de detectar los casos relevantes. Al realizar la validación de la sensibilidad se puede considerar como un algoritmo con sensibilidad alta que no se le escapan muchos positivos. Es decir, el modelo es capaz de identificar un 96.875% de los estudiantes que estarían interesados en continuar sus estudios de posgrados en la universidad como se puede observar en la Tabla 12 en donde se presentan las diferencias entre las métricas del modelo inicial y el modelo final con la selección de los hiperparámetros que optimizaron el modelo.

Tabla 12 - Comparativa de métricas con el modelo final de árboles de decisión

Métrica	*modelo inicial max_depth=None class_weight=None ccp_alpha=0.0*	*max_depth=4 class_weight=”balanced” ccp_alpha=0.0*	*modelo optimizado max_depth=9 class_weight='balanced' ccp_alpha=0.0242*
Accuracy	98,6854%	98,2159%	98,9671%
Precision	96,7948%	90,7514%	96,2732%
Recall	94,3750%	98,125%	96,8750%
F1	95,5696%	94,2942%	96,5732%

Otro aspecto que se tuvo en cuenta cuando se realizó la construcción del modelo fue determinar la importancia de los predictores en el modelo, calculado como la reducción total (normalizada) en el criterio de división, en este caso el índice Gini, que consigue el predictor en las divisiones en las que participa. Si un predictor no fue seleccionado en ninguna división, no fue incluido en el modelo y por lo tanto su importancia fue de 0. En la Tabla 13 se puede observar la importancia de los predictores en el modelo, en donde el predictor referente a las becas es el de mayor importancia, seguido de los créditos financieros, créditos académicos y la cantidad de semestres matriculados.

Tabla 13 - Importancia de los predictores

Predictor	Importancia
Becas/Descuentos	0.91
Créditos Financieros	0.06
Créditos Académicos	0.01
Semestres Matriculados	0.01
Promedio Académico	0.00
Monitorias	0.00
Ingeniería Civíl	0.00
Edad	0.00
Materias Perdidas	0.00
Consejerías	0.00
Lugar Nacimiento	0.00
Ingeniería Industrial	0.00
Doble Programa	0.00
Ingeniería de Sistemas	0.00
Ingeniería Electrónica	0.00
Biología	0.00
Movilidad	0.00
Beca Permanente	0.00
Distinciones	0.00
Género	0.00
Pruebas Académicas	0.00
Ciudad	0.00
Matemáticas Aplicadas	0.00

5.4.1.5. Bosques aleatorios. La matriz de confusión aplicando bosques aleatorios se observa en la Figura 23.

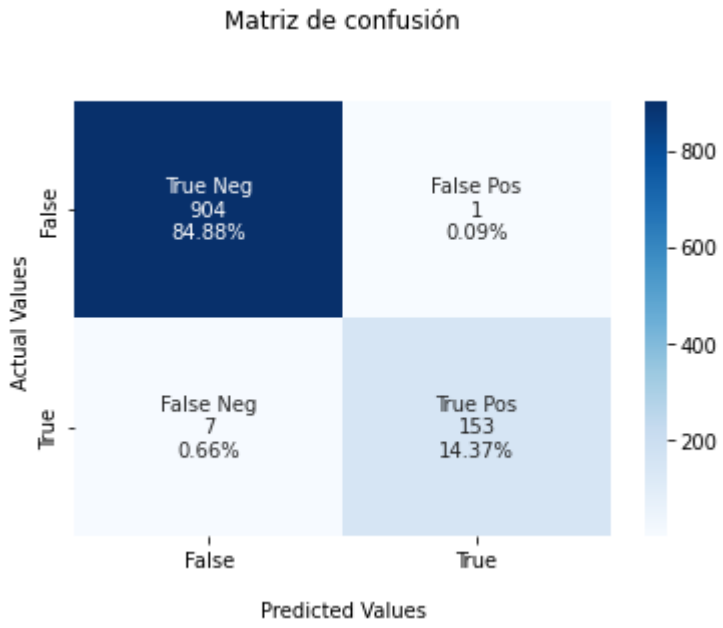


Fig. 23 Matriz de confusión de bosques aleatorios

El modelo final con la técnica de bosques aleatorios es capaz de identificar un 96.25% de los estudiantes que estarían interesados en continuar sus estudios de posgrados en la universidad como se puede observar en la Tabla 14 en donde se presentan las diferencias entre el modelo inicial y sus variantes hasta lograr obtener el modelo con los parámetros óptimos.

Tabla 14 - Comparativa de métricas con el modelo final de bosques aleatorios

Métrica	*modelo inicial criterion=gini max_depth=None max_features='auto' n_estimators=100 class_weight=None*	*criterion=gini max_depth=12 max_features='auto' n_estimators=150 class_weight='balanced'*	*modelo optimizado criterion=entropy max_depth=7 max_features=11 n_estimators=50 class_weight='balanced'*
Accuracy	99,2488%	99,061%	99,2488%
Precision	98,7179%	97,4883%	98,7179%
Recall	96,25%	96,25%	96,2500%
F1	97,4683%	96,8553%	97,4683%

Si bien es cierto que el proceso de los bosques aleatorios consigue mejorar la capacidad predictiva en comparación a los modelos basados en un único árbol, esto tiene un coste asociado, la

interpretabilidad del modelo se reduce. Al tratarse de una combinación de múltiples árboles, no es posible obtener una representación gráfica sencilla del modelo y no es inmediato identificar de forma visual qué predictores son más importantes. Sin embargo, se han desarrollado estrategias para cuantificar la importancia de los predictores que hacen de los modelos de bosques aleatorios una herramienta muy potente, no solo para predecir, sino también para el análisis exploratorio. Dos de estas medidas son: importancia por permutación e impureza de nodos. En la Tabla 15 y en la Tabla 16 se puede ver representada la importancia que tienen cada uno de los predictores en los modelos a partir de las diferentes técnicas mencionadas anteriormente. Cabe aclarar que la importancia de las características basadas en purezas puede inflar la importancia de las características numéricas. Además, la importancia de la característica basada en la pureza de los bosques aleatorios se ve afectada por el cálculo de las estadísticas derivadas del conjunto de datos de entrenamiento: la importancia puede ser alta incluso para las características que no son predictivas de la variable que se requiere predecir [21].

Tabla 15 - Importancia de los predictores por pureza de nodos

Predictor	Importancia
Becas/Descuentos	0.63
Créditos Financieros	0.30
Promedio Académico	0.02
Movilidad	0.01
Edad	0.01
Créditos Académicos	0.01
Materias Perdidas	0.01
Semestres Matriculados	0.01
Lugar Nacimiento	0.01
Consejerías	0.00
Monitorias	0.00
Ingeniería Civíl	0.00
Biología	0.00
Género	0.00
Distinciones	0.00
Ingeniería Industrial	0.00
Doble Programa	0.00
Ingeniería Electrónica	0.00
Matemáticas Aplicadas	0.00
Ciudad	0.00
Pruebas Académicas	0.00
Ingeniería de Sistemas	0.00
Beca Permanente	0.00

Tabla 16 - Importancia de los predictores por permutación

Predictor	Importancia (mean)	Importancia (std)
Becas/Descuentos	0.39	0.00
Créditos Financieros	0.33	0.00
Promedio Académico	0.05	0.00
Edad	0.03	0.00
Lugar Nacimiento	0.03	0.00
Créditos Académicos	0.03	0.01
Consejerías	0.02	0.00
Doble Programa	0.02	0.01
Materias Perdidas	0.02	0.01
Semestres Matriculados	0.01	0.01
Ingeniería Civíl	0.01	0.01
Ingeniería de Sistemas	0.00	0.00
Ingeniería Industrial	0.00	0.00
Ingeniería Electrónica	0.00	0.00
Biología	0.00	0.00
Movilidad	0.00	0.00
Beca Permanente	0.00	0.00
Monitorias	0.00	0.00
Distinciones	0.00	0.00
Género	0.00	0.00
Pruebas Académicas	0.00	0.00
Ciudad	0.00	0.00
Matemáticas Aplicadas	0.00	0.00

La Figura 24 muestra la importancia de los predictores en el modelo, a partir de ella se puede identificar las características de las becas/descuentos y créditos financieros como los predictores más influyentes, acorde a los datos de entrenamiento. Por otro lado, las características del promedio académico, edad, lugar de nacimiento, créditos académicos, consejerías, materias perdidas y doble programa indican cómo el modelo ha utilizado poco dichas características para la predicción. El resto de características al ser muy cercanas a 0, quiere decir que no son utilizadas por el modelo para la predicción.

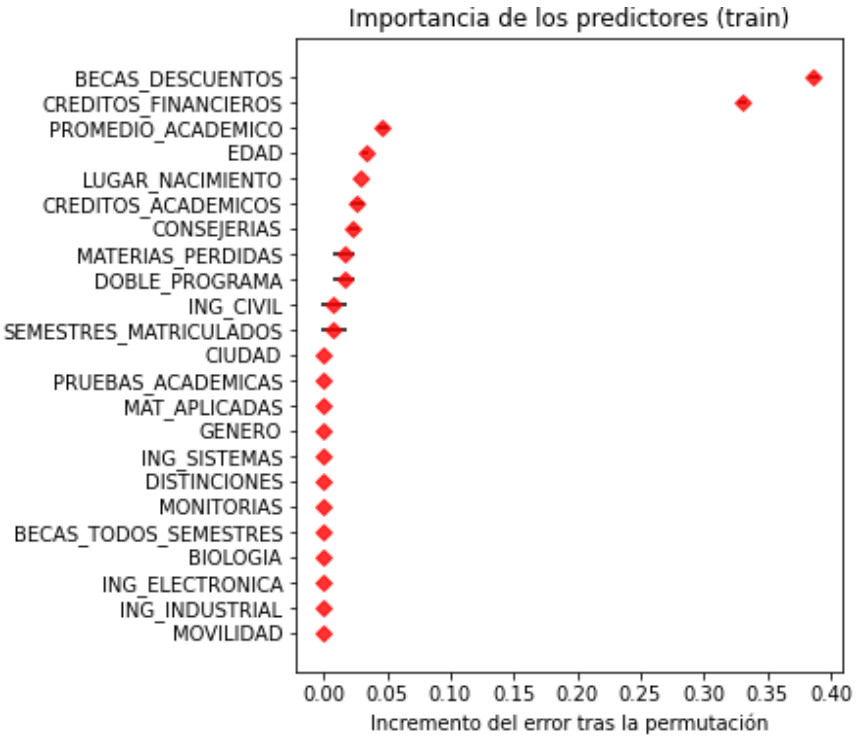


Fig. 24 Importancia de los predictores por permutación

5.4.1.6. Potenciación del gradiente. La matriz de confusión aplicando potenciación del gradiente se observa en la Figura 25.

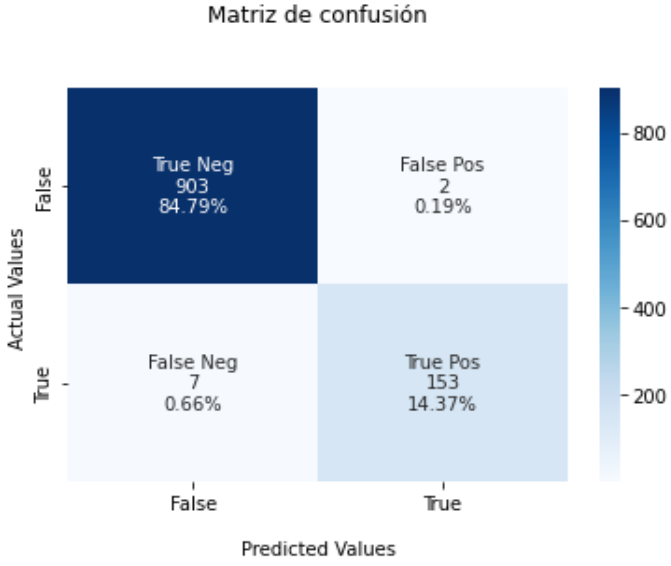


Fig. 25 Matriz de confusión de potenciación del gradiente

El modelo final con la técnica de potenciación del gradiente es capaz de identificar un 95.625% de los estudiantes que estarían interesados en continuar sus estudios de posgrados en la universidad como se puede observar en la Tabla 17 en donde se presentan las diferencias entre el modelo inicial y sus variantes hasta lograr obtener el modelo con los parámetros óptimos.

Tabla 17 - Comparativa de métricas con el modelo final de bosques aleatorios

Métrica	*modelo inicial learning_rate=0.1 max_depth=3 max_features=None n_estimators=100 subsample=1.0*	*modelo optimizado learning_rate=0.01 max_depth=3 max_features='auto' n_estimators=1000 subsample=0.5*
Accuracy	99,1549%	99,1549%
Precision	98,7096%	98,7096%
Recall	95,625%	95,625%
F1	97,1428%	97,1428%

Al evaluar y considerar la importancia de los predictores (influencia que tiene cada predictor sobre una determinada métrica de evaluación del modelo) se pueden identificar aquellas características que contribuyen en mayor y menor medida al modelo. Si el predictor permutado contribuye al modelo, es de esperar que el modelo aumente su error, ya que se pierde la información que proporcionaba esa variable. Sin embargo, cabe tomar algunas precauciones en su interpretación. Lo que cuantifican los predictores es la influencia que tienen sobre el modelo, no su relación con la variable respuesta.

Por otro lado, la importancia de los predictores por pureza de nodos cuantifica el incremento total en la pureza de los nodos debido a divisiones en las que participa el predictor (promedio de todos los árboles). En la Tabla 18 y en la Tabla 19 se aprecia la importancia de los predictores aplicando las diferentes estratégicas.

Tabla 18 - Importancia de los predictores por pureza de nodos (potenciación del gradiente)

Predictor	Importancia
Becas/Descuentos	0.85
Créditos Financieros	0.09
Edad	0.01
Promedio Académico	0.01
Créditos Académicos	0.01

Lugar Nacimiento	0.00
Semestres Matriculados	0.00
Materias Perdidas	0.00
Doble Programa	0.00
Distinciones	0.00
Pruebas Académicas	0.00
Matemáticas Aplicadas	0.00
Consejerías	0.00
Movilidad	0.00
Ingeniería Industrial	0.00
Género	0.00
Ingeniería Civíl	0.00
Monitorias	0.00
Ingeniería Electrónica	0.00
Biología	0.00
Ingeniería de Sistemas	0.00
Beca Permanente	0.00
Ciudad	0.00

Tabla 19 - Importancia de los predictores por permutación (potenciación del gradiente)

Predictor	Importancia (mean)	Importancia (std)
Becas/Descuentos	0.39	0.00
Créditos Financieros	0.34	0.00
Edad	0.04	0.01
Promedio Académico	0.04	0.00
Créditos Académicos	0.03	0.00
Doble Programa	0.03	0.00
Lugar Nacimiento	0.03	0.00
Pruebas Académicas	0.03	0.00
Materias Perdidas	0.02	0.00
Matemáticas Aplicadas	0.02	0.00
Semestres Matriculados	0.02	0.00
Distinciones	0.01	0.01
Género	0.00	0.00
Consejerías	0.00	0.00
Monitorias	0.00	0.00
Beca Permanente	0.00	0.00

Ciudad	0.00	0.00
Biología	0.00	0.00
Ingeniería Civíl	0.00	0.00
Ingeniería Electrónica	0.00	0.00
Ingeniería Industrial	0.00	0.00
Ingeniería de Sistemas	0.00	0.00
Movilidad	0.00	0.00

La Figura 26 muestra la importancia de los predictores en el modelo, a partir de ella se puede identificar las características de las becas/descuentos y créditos financieros como los predictores más influyentes, acorde a los datos de entrenamiento.

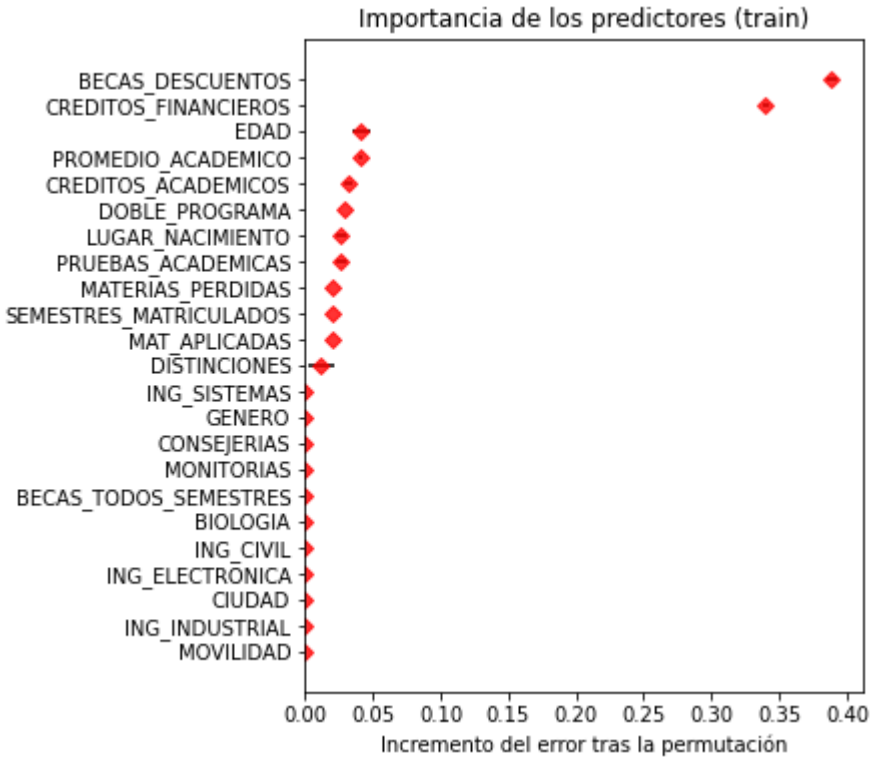


Fig. 26 Importancia de los predictores por permutación (potenciación del gradiente)

5.5. FASE V: EVALUACIÓN

En esta fase se evaluaron los resultados aplicando las principales métricas de evaluación del rendimiento usadas en machine learning, para seleccionar el mejor modelo a fin de poder determinar si un estudiante egresado de pregrado continuará su formación académica en posgrados en la universidad.

5.5.1 EVALUACIÓN DE RESULTADOS. Desde el punto de vista objetivo, se efectuó el análisis con los indicadores estadísticos que se obtuvieron al ejecutar los modelos con los hiperparámetros optimizados, con el objetivo de determinar si el resultado de dicha evaluación de cada modelo puede predecir con un grado de alta confiabilidad si un estudiante de pregrado continuaría sus estudios en posgrados en la universidad. A continuación, se encuentra el análisis de cada modelo.

5.5.1.1. Árboles de decisión. Desde el punto de vista de exactitud del modelo de árboles de decisión, aunque se obtuvo un porcentaje de 98,9671% de estudiantes clasificados correctamente, el tener un desbalance notable entre las clases de variables de destino en los datos, es una desventaja significativa porque se corre el riesgo de siempre predecir que los estudiantes no continuarán sus estudios en posgrados, teniendo en cuenta que la clase mayoritaria en el conjunto de datos correspondía a la no continuidad. Precisamente, al tener un dataset desbalanceado es aconsejable el indicador F1 (en el caso de los árboles de decisión el resultado fue de 96,5732%) si se tiene una distribución de clases desigual. La sensibilidad del modelo, definida como la capacidad de predecir un resultado positivo cuando el resultado real es positivo, dio como resultado un 96,8750% y la precisión, definida como el número de elementos identificados correctamente como positivos de un total de elementos identificados como positivos, dio como resultado un 96,2732%.

5.5.1.2. Bosques aleatorios. Al igual que sucedió con los árboles de decisión, en los bosques aleatorios se obtuvo un porcentaje alto de precisión del modelo (99,2488%) de estudiantes clasificados correctamente. En el caso de la métrica de F1, para los bosques aleatorios el resultado del indicador fue de 97,4683%. La sensibilidad del modelo dio como resultado un 96,25% y la precisión dio como resultado un 98,7179% lo cual es mayor al valor obtenido en los árboles de decisión. Esto significa que un 98% de los registros logran ser clasificados o identificados correctamente como positivos.

5.5.1.3. Potenciación del gradiente. Al igual que sucedió con los árboles de decisión y los bosques aleatorios, en la potenciación del gradiente se obtuvo un porcentaje alto de precisión del modelo (99,1549%) de estudiantes clasificados correctamente. En el caso de la métrica de F1 el resultado del indicador fue de 97,1428%. La sensibilidad del modelo dio como resultado un

95,625% y la precisión dio como resultado un 98,7096% lo cual es mayor al valor obtenido en los árboles de decisión y bosques aleatorios.

5.5.2 COMPARATIVA DEL DESEMPEÑO ENTRE MODELOS. Después de realizar los diferentes experimentos de modelado con los algoritmos de clasificación seleccionados, para los cuales se combinaron técnicas de tuning de parámetros individuales, así como búsqueda en cuadrícula de hiperparámetros con validación cruzada; comparando el desempeño de los modelos en términos de las métricas de F1, precisión y recall, se presenta en la Tabla 20 los mejores resultados obtenidos de cada algoritmo con sus métricas de desempeño.

Tabla 20 - Comparativo del desempeño entre modelos

Modelo	F1-Score Macro	Recall Macro	Precision Macro	Accuracy
Decision Tree	0.966	0.969	0.963	0.990
Random Forest	0.972	0.963	0.981	0.992
Gradient Boosting	0.971	0.956	0.956	0.992

Partiendo que la métrica F1-Score es la medida de desempeño seleccionada con el propósito de comparar los modelos evaluados, teniendo en cuenta principalmente el desbalance presentado en las clases, la cual busca minimizar los falsos negativos, sin olvidarse de los costos asociados a los falsos positivos, se observa en la Tabla 20 que los mejores resultados se obtienen con Random Forest que, adicionalmente, obtuvo una precisión que supera a los modelos de Decision Tree y Gradient Boosting. Por otro lado, es importante mencionar que con Random Forest se logra reducir la varianza de los árboles individuales al combinar la aleatoriedad (cada árbol entrenado con un subset diferente) y la agregación (al agregar los resultados para generar la predicción, los árboles que no funcionaron tan bien no tuvieron un impacto significativo en el resultado final). También se seleccionó Random Forest por encima de Gradient Boosting porque este último presenta desventajas en la escalabilidad debido a la naturaleza secuencial de impulsar, dificultando la tarea de paralelizar.

5.6. FASE VI: INFORME

En esta fase se produjo un informe final que resume los resultados obtenidos en la aplicación del modelo y en la utilización de la metodología. Así como también se hizo una revisión general del

proyecto con aquellos puntos que presentaron dificultad y la identificación de lecciones aprendidas.

5.6.1 INFORME FINAL. La aplicación de la metodología CRISP-DM en este proyecto permitió encontrar un comportamiento predictivo a la hora de determinar la continuidad académica de estudiantes de pregrado de la Facultad de Ingeniería y Ciencias de la Pontificia Universidad Javeriana. La etapa de la preparación de los datos fue la más extensa y compleja de todo el proceso por la alta inconsistencia presentada en algunos datos mencionados durante el proyecto que se encontraban por fuera de lo establecido en el Reglamento del Estudiante de la universidad, lo que implicó el análisis detenido del reglamento, sus excepciones e investigación a nivel interno de los procesos que enmarcan el ámbito académico, administrativo y disciplinario. También es importante mencionar la limpieza realizada previa a la preparación de los datos, lo que generó un proceso adicional dentro de la aplicación de la metodología que puede evitarse en futuros trabajos al ejecutar procesos exhaustivos de calidad de datos dentro de los sistemas fuente de la información de los trabajos futuros.

Una vez se realizó la preparación y limpieza de los datos, se generaron los modelos de predicción aplicando las técnicas de clasificación mencionadas en el proyecto, que determinó la adecuación de los modelos generados determinando su pertinencia y validez al ser lo suficientemente fiables. En este caso en particular del proyecto, se determina que uno de los tres modelos generados es el que genera un mayor índice de confianza y es el recomendado para su aplicación sobre futuros conjuntos de datos a predecir. Sin embargo, es importante mencionar que los 3 modelos generados representan diferencias muy pequeñas en cuanto a los resultados obtenidos por lo que podemos considerar que la aplicación de cualquiera de los 3 modelos es válida teniendo en cuenta los buenos resultados.

Por otro lado, la evaluación de los modelos permitió encontrar aquellas características más importantes para el modelo en su predicción. Al realizar un análisis de los datos, se logró determinar que las becas representan una característica muy influyente para que un estudiante tome la decisión de continuar o no sus estudios en posgrados en la universidad, lo que representa información muy valiosa para la Oficina de Promoción a la hora de realizar procesos de atracción para los egresados de la universidad.

6. CONCLUSIONES

Las etapas de entendimiento y preparación de los datos aplicando la metodología CRISP-DM son las más demandantes dentro de cualquier proyecto de minería de datos y este proyecto no fue la excepción. Estas etapas permitieron tener un amplio panorama de los datos que sirvieron para determinar diferentes aspectos relevantes dentro de los procesos académicos de los estudiantes cuando se encuentran en su etapa de estudios de pregrado y aspectos administrativos de la universidad dentro del acompañamiento integral a la formación académica de los estudiantes.

Cuando se comprende de manera integral la dinámica de los procesos educativos en instituciones de educación superior (IES) como lo es la Pontificia Universidad Javeriana, se logra acotar los atributos que se tienen a consideración al momento de generar modelos de predicción, es decir, al tener un conocimiento del funcionamiento de dichos procesos académicos y administrativos, permite agilizar la comprensión que se debe tener en cuenta en la etapa de entendimiento de los datos, generando sentido a los atributos seleccionados y descartando aquellos que pueden llegar a ser considerados como irrelevantes para los modelos al no contener información de importancia y no aportar valor al análisis, teniendo en cuenta que, por ejemplo, para el caso del tipo de identificación del estudiante, solo se presentaban 3 categorías (cédula de ciudadanía, cédula de extranjería y pasaporte), pero solo estaba presente 1 registro para la categoría de cédula de extranjería y pasaporte respectivamente. De igual manera se descartó el atributo ciudad, teniendo en cuenta que el estudiante durante su etapa académica, normalmente su lugar de residencia corresponde a la ciudad de Cali o sus alrededores.

Sin embargo, es de vital importancia realizar una profunda evaluación y análisis crítico de cada actividad realizada en estas etapas, teniendo en cuenta que pasar desapercibidos ciertos detalles generan reprocesos que impactaron en el tiempo de desarrollo del proyecto y cambios no previstos. Por lo anterior, fue un enorme reto la dedicación minuciosa sobre cada aspecto y cada atributo presente en el conjunto de datos, a partir de la generación de las diferentes gráficas y estadísticas que permitieron identificar diferentes aspectos que pudieran generar ruido en los modelos o inconsistencias en la información del conjunto de datos.

En cuanto al modelado, es importante seleccionar las técnicas más apropiadas para el proyecto considerando el objetivo principal del mismo y la relación con las herramientas de minería de datos existentes. Esta selección permitió hacer una interpretación de los resultados de los modelos de aprendizaje y obtener los atributos que están más relacionados con la decisión del estudiante de continuar o no con los estudios de posgrado.

Desde el punto de vista de los resultados obtenidos, es importante mencionar que son producto del análisis detallado y minucioso de resultados cercanos en las etapas previas, en particular, el análisis y limpieza de datos, teniendo en cuenta el desbalance de clases a predecir como resultado de la naturaleza de los datos, se logró generar modelos con alta confiabilidad teniendo en cuenta este desbalance presentado en los datos; se esperaba obtener un balance de los datos en una proporción de 80-20 de egresados que no continúan sus estudios en posgrados y estudiantes que si lo hacen, pero se logró obtener una proporción de 85-15 teniendo en cuenta que pueden interferir muchos factores como la oferta académica, la condición económica, etc.

Finalmente, aunque se obtuvieron muy buenos resultados en el entrenamiento y evaluación inicial de los modelos, se optimizaron sus parámetros producto del ejercicio práctico como objetivo y fin de la realización del presente trabajo, que aunque permitió obtener mejores resultados, sirvió de aprendizaje para futuros trabajos de minería de datos en el campo de predicción de continuidad académica de la universidad, teniendo en cuenta que el conjunto de datos trabajado fue filtrado según la disponibilidad con la cual se contaba de los datos en los sistemas de información.

También es importante destacar el hallazgo de atributos como lo es la obtención de becas para los estudiantes pues define un comportamiento importante frente a la continuidad o no para estudios de posgrados en la universidad, convirtiéndose así en un atributo que puede ser fundamental para la perfilación de egresados por parte de la Oficina de Promoción de la universidad.

7. TRABAJOS FUTUROS

Como continuación del presente trabajo en un futuro es importante que se pueda generar un modelo para las otras facultades de la universidad, así como también poder tener características referentes a los planes de estudio de las diferentes carreras. También sería importante contar con la recopilación y análisis de la información laboral del egresado en un periodo de tiempo determinado para conocer y evaluar sus condiciones sociodemográficas, económicas y sociales que permitan tener un panorama más amplio de características que permitan predecir la continuidad de un egresado a programas de posgrados. Finalmente, otro aspecto que se puede tener en consideración es el referente a la participación del egresado en los diferentes programas de educación continua que brinda la universidad, en donde se podría generar información mucho más actualizada del egresado.

8. BIBLIOGRAFÍA

[1] Gobernación de Antioquia *et al.*, "OBSERVATORIO CT + i," *Cons. Nac. Política Económica Y Soc.*, vol. 53, no. 1, pp. 34–55, 2015.

[2] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1432–1462, 2014, doi: 10.1016/j.eswa.2013.08.042.

[3] S. Basu, "Data Mining," *Georgia State University Fall*, 1997. http://www6.uniovi.es/hypvis/applicat/data_mining/data_mining.html (accessed Apr. 05, 2021).

[4] M. Jose, P. S. Kurian, and V. Biju, "Progression analysis of students in a higher education institution using big data open source predictive modeling tool," *2016 3rd MEC Int. Conf. Big Data Smart City, ICBDSC 2016*, pp. 113–117, 2016, doi: 10.1109/ICBDSC.2016.7460352.

[5] Z. Bošnjak, O. Grljević, and S. Bošnjak, "CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data," *Proc. - 2009 5th Int. Symp. Appl. Comput. Intell. Informatics, SACI 2009*, no. 114, pp. 509–514, 2009, doi: 10.1109/SACI.2009.5136302.

[6] J. W. Seifert, "CRS Report for Congress Data Mining," *Reading*, pp. 1–16, 2004.

[7] X. Yu and S. Wu, "Typical Applications of Big Data in Education," *Proc. - 2015 Int. Conf. Educ. Innov. Through Technol. EITT 2015*, pp. 103–106, 2016, doi: 10.1109/EITT.2015.29.

[8] S. Yu, D. Yang, and X. Feng, "A Big Data Analysis Method for Online Education," *Proc. - 10th Int. Conf. Intell. Comput. Technol. Autom. ICICTA 2017*, vol. 2017-Octob, pp. 291–294, 2017, doi: 10.1109/ICICTA.2017.71.

[9] A. Jamil, M. Abdullah, M. A. Javed, and M. S. Hassan, "Comprehensive Review of Challenges Technologies for Big Data Analytics," *2018 IEEE Int. Conf. Comput. Commun. Eng. Technol. CCET 2018*, pp. 229–233, 2018, doi: 10.1109/CCET.2018.8542219.

[10] A. Anjewierden, H. Gijlers, N. Saab, and R. De Hoog, "Brick: Mining pedagogically interesting sequential patterns," *EDM 2011 - Proc. 4th Int. Conf. Educ. Data Min.*, pp. 341–342, 2011.

[11] C. Russo, H. Ramón, N. Alonso, B. Cicerchia, L. Esnaola, and J. P. Tessore, "Tratamiento Masivo de Datos Utilizando Técnicas de Machine Learning," *XVIII Work. Investig. en Ciencias la Comput. (WICC 2016, Entre Ríos, Argentina)*, p. 131, 2016.

[12] A. González, *¿Qué es Machine Learning?* 2014, p. 14.

[13] J. Villena Román, "CRISP-DM: La metodología para poner orden en los proyectos," *Sngular*, 2016. https://www.sngular.com/es/data-science-crisp-dm-metodologia/ (accessed Apr. 18, 2021).

[14] P. C. Ncr *et al.*, "Step-by-step data mining guide," *SPSS inc*, vol. 78, pp. 1–78, 2000, [Online]. Available: http://www.crisp-dm.org/CRISPWP-0800.pdf.

[15] J. F. Vallalta, "CRISP-DM: una metodología para minería de datos en salud - healthdataminer.com." https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/ (accessed Apr. 17, 2021).

[16] B. Herold, “The future of big data and analytics in K-12 education,” *Ed Week*, 2016. https://www.edweek.org/policy-politics/the-future-of-big-data-and-analytics-in-k-12-education/2016/01 (accessed Mar. 20, 2021).

[17] A. I. Oviedo Carrascal and J. Jiménez Giraldo, “Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO,” *Rev. Politécnica*, vol. 15, no. 29, pp. 128–140, 2019, doi: 10.33571/rpolitec.v15n29a10.

[18] P. Vats and K. Samdani, "Study on Machine Learning Techniques In Financial Markets", 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2019, pp. 1-5, doi: 10.1109/ICSCAN.2019.8878741.

[19] H. I. Bulbul and Ö. Unsal, "Comparison of Classification Techniques used in Machine Learning as Applied on Vocational Guidance Data," 2011 10th International Conference on Machine Learning and Applications and Workshops, 2011, pp. 298-301, doi: 10.1109/ICMLA.2011.49.

[20] A. Muller and S. Guido, “Introduction to Machine Learning with Python”, vol 1, O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, 2006, pp. 80-92, doi: 10.1109/ICICTA.2016.30.

[21] K. Johnson and M. Kuhn, “Applied Predictive Modeling”, vol 1, Springer, New York, 2013, pp. 202-203, doi: 10.1109/ICICTA.2013.78.

[22] Gomes Mantovani, R., Horváth, T., Cerri, R., Barbon Junior, S., Vanschoren, J., and Ferreira de Carvalho, A. C. P. de L., “An empirical study on hyperparameter tuning of decision trees”, arXiv e-prints, 2018.

[23] L. Villalobos and C. Quesada, “Comparative study of random search hyper-parameter tuning for software effort estimation”, Proceedings of the 17th International Conference on Predictive Models and Data Analytics in Software Engineering, co-located with ESEC/FSE 2021, doi: 10.1145/3475960.3475986